Approximation properties of Sum-Up Rounding

Von der
Carl-Friedrich-Gauß-Fakultät
der Technischen Universität Carolo-Wilhelmina zu Braunschweig

zur Erlangung des Grades eines
Doktors der Naturwissenschaften (Dr. rer. nat.)

genehmigte Dissertation

von
Paul Manns
geboren am 17. Februar 1990
in Ulm

Eingereicht am: 18. Juli 2019
Disputation am: 29. Oktober 2019
1. Referent: Prof. Dr. Christian Kirches
2. Referent: Prof. Dr. Christian Meyer

2019

Abstract

Optimization problems that involve discrete variables are exposed to the conflict between being a powerful modeling tool and often being hard to solve. Infinite-dimensional processes, as e.g. described by differential equations, underlying the optimization may lead to the need to solve for distributed discrete control variables.

This work analyzes approximation arguments that replace the need for solving the optimization problem by the need for first solving a relaxation and second computing appropriate roundings to regain discrete controls. We provide sufficient conditions on rounding algorithms and their grid refinement strategies that allow to prove approximation of the relaxed controls by the discrete controls in weaker topologies, a feature due to the infinite-dimensional vantage point. If the control-to-state mapping of the underlying process exhibits suitable compactness properties, state vector approximation follows in the norm topology as well as, under additional assumptions, optimality principles of the computed discrete controls. The conditions are verified for representatives of the family of Sum-Up Rounding algorithms.

We apply the arguments on different classes of mixed-integer optimization problems that are constrained by partial differential equations. Specifically, we consider discrete control inputs, which are distributed in the time domain, for evolution equations that are governed by a differential operator that generates a strongly continuous semigroup, discrete control inputs, which are distributed in multi-dimensional spatial domains, for elliptic boundary value problems and discrete control inputs, which are distributed in space-time cylinders, for evolution equations that are governed by differential operators such that the corresponding Cauchy problem satisfies maximal parabolic regularity. Furthermore, we apply the arguments outside the scope of partial differential equations to a signal reconstruction problem. Computational results illustrate the findings.

Zusammenfassung

Optimierungsprobleme mit diskreten Variablen befinden sich im Spannungsfeld zwischen hoher Modellierungsmächtigkeit und oft schwerer Lösbarkeit. Zur Optimierung unendlichdimensionaler Prozesse, z.B. beschrieben mit Hilfe von Differentialgleichungen, kann die Lösung nach verteilten diskreten Kontrollvariablen erforderlich sein.

Diese Arbeit untersucht Approximationsargumente, mit deren Hilfe die Notwendigkeit einer Lösung des Optimierungsproblems durch die Notwendigkeit zuerst eine Relaxierung zu lösen und anschließend eine passende Rundung zu berechnen, um wieder diskrete Kontrollvariablen zu erhalten, ersetzt wird. Wir geben hinreichende Bedingungen an Rundungsalgorithmen und ihre Gitterverfeinerungsstrategien an, um eine Approximation der relaxierten Kontrollvariablen mit den diskreten Kontrollvariablen in schwächeren Topologien zu erhalten, was aus der unendlichdimensionalen Betrachtung des Problems folgt. Falls der Steuerungs-Zustands-Operator des zugrundeliegenden Prozesses passende Kompaktheitseigenschaften aufweist, folgen die Approximation der Zustandsvektoren in der Normtopologie und, unter zusätzlichen Bedingungen, Optimalitätsprinzipien für die berechneten diskreten Kontrollvariablen. Die Bedingungen werden für Repräsentanten der Familie von Sum-Up Rounding Algorithmen nachgewiesen.

Wir wenden die Argumente auf verschiedene Klassen von gemischt-ganzzahligen Optimierungsproblemen, die von partiellen Differentialgleichungen beschränkt werden, an. Insbesondere betrachten wir diskrete, in der Zeit verteilte, Steuerungen in Evolutionsgleichungen mit Differentialoperatoren, die stark stetige Halbgruppen erzeugen; diskrete, mehrdimensional im Ort verteilte, Steuerungen in elliptischen Randwertproblemen und diskrete, in Ort und Zeit verteilte, Steuerungen in Evolutionsgleichungen mit Differentialoperatoren, deren zugehörige Cauchyprobleme maximale parabolische Regularität aufweisen. Des Weiteren wenden wir die Argumente außerhalb des Kontexts partieller Differentialgleichungen auf ein Signalrekonstruktionsproblem an. Numerische Beispiele illustrieren die gezeigten Resultate.

Contents

Chapter 1

Introduction

> Although this may seem as a paradox, all exact science is dominated by the idea of approximation.
>
> (Bertrand Russell — [95, p. 65])

This holds especially true for mathematics, an exact science if any such thing exists, where approximation techniques lie at the heart of many different fields, in particular analysis and numerics, the latter providing means to actually compute approximate solutions to mathematical problems. Mixed-Integer Optimal Control Problems (MIOCPs) constitute a rich class of mathematical problems that can be employed to model many *real-world* applications ranging from topology optimization, see e.g. [52], over optimum experimental design, see e.g. [99], and energy management of buildings, see e.g. [122], to automobile test drives, see e.g. [41].

Approximating solutions of an MIOCP with a *first discretize, then optimize* strategy yields a Mixed-Integer Nonlinear Program (MINLP), which may exhibit a cataclysmic computational demand when attacking it with the established techniques for mixed integer optimization, namely branch and bound and its wealth of variants, which date back at least to the 1960s, cf. [25,77]. From an abstract point of view, this is unsurprising as MINLPs are known to be NP-complete, see [40]. We note as in [80] that, from a computational point of view, the difficulties are due to the fatal combination of the high number of variables arising from the discretization of the state equation of the controlled model with the curse of dimensionality arising from the decision tree of the integer (control) variables.

This drives the search for approximation techniques that are capable of producing points of low objective value and low infeasibility at low computational cost. In this context, this work analyzes a chain of approximation arguments for a relaxation-based procedure, in which roundings are computed from the solution of a continuous relaxation in linear time w.r.t. the mesh size of the rounding grid. The procedure is advantageous if the relaxation can be assumed to be solvable reasonably fast.

Although this may seem as a paradox, the infinite-dimensional point of view enables us to find powerful approximation properties that are inaccessible in the finite-dimensional setting and in some sense even simplify the problem. In particular, we prove that the family of Sum-Up Rounding (SUR) algorithms computes integer-valued approximants of relaxed controls in weaker topologies of L^p spaces. This in turn leverages compactness properties of the solution mappings of the underlying state equations. The peculiarity lies with the fact that the compactness induces that the approximants of the state vector converge to the (optimal) state vector of the relaxation in norm. This passes the approximation on to norm-continuous mappings, e.g. the objective of the MIOCP. The approximation quality can be controlled by the mesh size of the rounding grid.

In some sense, the methodology replaces the need to solve a potentially NP-hard or unsolvable problem by the need to solve a relaxation and compute the rounding on sufficiently fine grids to achieve a desired accuracy in terms of feasibility and optimality. The results suggest that minimizing integer controls of MIOCPs in L^p spaces might not exist in general and the existence of an approximating sequence might be the best we can hope for.

Our findings complement the rich theory of bang-bang-type approximation properties, in particular the Lyapunov convexity theorem and the Filippov–Ważewski theorem, in two ways. First, they can be interpreted as an extension and application to problems involving integer variables. Second, we provide an algorithmic framework that satisfies the assumptions needed for the analysis to apply on the one hand and is of constructive and implementation-friendly algorithmic nature on the other hand.

1.1 Problem statement

This work investigates MIOCPs of the following form:

$$\begin{aligned} \min_{y,u,\boldsymbol{v}} \; & J(y,u,\boldsymbol{v}) \\ \text{s.t. } & y = S(u,\boldsymbol{v}) \\ & 0 \leq c(y(t,x),u(t,x),\boldsymbol{v}(t,x)) \text{ a.e.} \\ & \boldsymbol{v}(t,x) \in \{v_1,\dots,v_M\} \text{ a.e.} \end{aligned} \tag{MIOCP}$$

Here, y denotes the state vector of an underlying process, u denotes a continuously-valued control vector and $\boldsymbol{v}$ denotes a discrete-valued control vector with finite codomain. These vectors are elements of function spaces and are defined on appropriate temporal or spatial domains or space-time cylinders depending on the problem formulation. The symbol S denotes the solution or control-to-state operator of the underlying process, often a differential equation. Its properties, in particular compactness, or complete continuity to be more precise, are of crucial importance in the remainder. The symbol c denotes a constraint function, which has a pointwise almost everywhere interpretation

by means of the corresponding superposition operator. We often refer to it as *mixed constraint* as it depends on the state, the continuous and the discrete control vector.

We employ Sager's *partial outer convexification* technique, see [98], to derive equivalent reformulations of (MIOCP) instances. To this end, the discrete-valued control variables are substituted by binary-valued ones satisfying a *one-hot encoding* or Special Ordered Set of Type 1 (SOS1) property. This gives rise to two relaxations. First, a *continuous relaxation* arises if $[0,1]$-valued controls replace the binary controls. Second, a *feasibility relaxation* arises by changing the pointwise a.e. mixed constraint $0 \leq c(y,u,\boldsymbol{v})$ a.e. to $-\delta \leq c(y,u,\boldsymbol{v})$ a.e. for some $\delta > 0$. A corresponding binary control satisfying the relaxed mixed constraint for some $\delta > 0$ is denoted by the symbol $\boldsymbol{\omega}^\delta$.

Both relaxations have different advantages and disadvantages, which are complementary to each other. Solutions for the continuous relaxation can be obtained without the problems imposed by discrete variables if the tools from nonlinear optimization can be applied to it. However, after having computed a solution for the continuous relaxation, there is no obvious means to relate it back to solutions of the original problem. On the other hand, the feasibility relaxation by some $\delta > 0$ does not necessarily give a simpler problem than the original one from the integer programming point of view. However, it can be related to the original problem in terms of feasibility quite easily: feasible points of the feasibility relaxation are at most δ-infeasible in the mixed constraint.

Sum-Up Rounding (SUR) is a class of algorithms that bridge the gap between solutions of the continuous relaxation, (RC) in the remainder, and feasible points of the feasibility relaxation, (BC_δ) in the remainder. It operates on a discretization grid whose mesh size can be used to control δ. Fortunately, the algorithm can be designed such that refining the discretization grid decreases δ. The particular choice of the rounding algorithm is a degree of freedom in the described methodology and alternative algorithms could be considered as well. Potential alternatives are put on the record in Chapters 2 and 4. To do justice to this fact, we strictly separate the rounding algorithms and their approximation properties from the analysis of the solution operators that exploit these approximation properties to establish approximation properties of the corresponding solutions of the state equations that constrain (MIOCP).

A large portion of this work is dedicated to the relationships of the optimization problems sketched above for various instances of (MIOCP). The relationships are visualized in Fig. 1.1. Especially the family of SUR algorithms and the *Optimality and Feasibility* of the rounded solution $(y^\delta, u^*, \boldsymbol{\omega}^\delta)$ for the feasibility relaxation are subject to investigation in the remainder.

1.2 Contribution

Let $\boldsymbol{\alpha}$ be a feasible $[0,1]$-valued control of the aforementioned continuous relaxation and $\boldsymbol{\omega}^{(h)}$ be a corresponding binary-valued rounding generated by a SUR algorithm on a grid

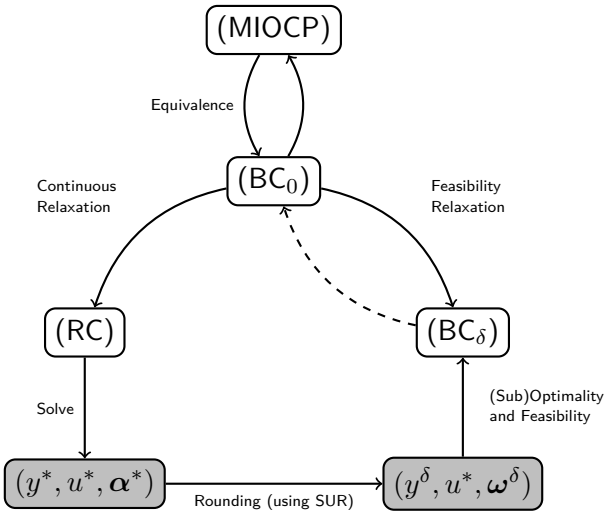


Figure 1.1: Investigated relationship between (MIOCP), (BC$_\delta$) and (RC).

with mesh size h. We prove the chain of arguments

$$h \to 0 \quad \underset{\text{(a)}}{\Rightarrow} \quad d^{(h)}(\boldsymbol{\omega}^{(h)}, \boldsymbol{\alpha}) \to 0 \quad \underset{\text{(b)}}{\Rightarrow} \quad \boldsymbol{\omega}^{(h)} \rightharpoonup \boldsymbol{\alpha} \quad \underset{\text{(c)}}{\Rightarrow} \quad y(\boldsymbol{\omega}^{(h)}) \to y(\boldsymbol{\alpha}), \tag{1.1}$$

in which the $d^{(h)}$ denote suitable pseudo-metrics arising from a sequence of grids on which the roundings are computed by virtue of the SUR algorithms. We note that our term mesh size will refer to the grid cells' volumes instead of the diameters. Regarding the implication (a), we prove a bound of the form $d^{(h)}(\boldsymbol{\omega}^{(h)}, \boldsymbol{\alpha}) \leq Ch$ with $C > 0$ being asymptotically tight for $M \to \infty$ for a SUR variant that also implies

$$d^{(h)}(\boldsymbol{\omega}^{(h)}, \boldsymbol{\alpha}) \to 0 \quad \Rightarrow \quad 0 \leq \liminf c(y^{(h)}, u, \boldsymbol{v}(\boldsymbol{\omega}^{(h)}))$$

for $\boldsymbol{v}(\boldsymbol{\omega}^{(h)})$ denoting the discrete control reconstructed from the rounding $\boldsymbol{\omega}^{(h)}$. In particular, we have $\boldsymbol{\omega}^\delta := \boldsymbol{\omega}^{(h)}$ with $\delta \to 0$ in the notation above. In the absence of such constraints, the constant $C > 0$ can be improved if the original SUR variant is used, see [68]. We provide an alternative proof for an asymptotically tight bound in this case. Clearly, the implications (b) and (c) are indifferent to the choice of the rounding algorithm. Regarding the implication (b), we prove that the $d^{(h)}(\boldsymbol{\omega}^{(h)}, \boldsymbol{\alpha}) \to 0$ induces convergence in the weak and weak* topologies of the L^p spaces and show how to extend the algorithms that have been developed for one-dimensional discrete variables to a multi-dimensional setting with discrete variables distributed in space and/or time. As a consequence of (b), convergence of the state vectors in the norm topology can be obtained if the solution operator of the state equation exhibits certain compactness properties. Therefore, we show sufficient compactness properties for several rich problem classes and apply them to obtain (c). The gained insights allow to apply SUR not only for processes described by differential equations as completely continuous or compact

solution operators can be found elsewhere as well. We show the applicability to a class of signal-reconstruction problems. Additionally to (a), (b) and (c), we

(d) prove consequences regarding the (sub)optimality of the computed sequences, $(\boldsymbol{\omega}^{(h)})_h$ and $(y(\boldsymbol{\omega}^{(h)}))_h$, for the feasibility relaxation and the original problem if $y(\boldsymbol{\alpha})$ and $\boldsymbol{\alpha}$ minimize the continuous relaxation, and

(e) provide computational results that demonstrate and validate our findings.

Publications Parts of the results for contribution (a), in particular results presented in Sections 6.4 and 6.5, have been submitted for publication in the article [82], which is currently under review. Parts of the results for contributions (b), (c), (d) and (e) in the one-dimensional case, in particular results in Sections 7.1, 9.1 and 10.1 have been submitted for publication in the article [80], which has been accepted for publication and is in the production process. Parts of the results for contributions (a), (b), (c), (d) and (e) in the multi-dimensional case, in particular results in Chapters 4 and 5 and Sections 7.2, 9.2 and 10.3, have been submitted for publication in the article [79], which is currently under review.

1.3 Structure of the remainder

We continue with a brief section that introduces the notation for the remainder of this work. The first part provides background information. Chapter 2 provides formal definitions of the partial outer convexification reformulation, the induced relaxations and the rounding algorithm as used in this work and embeds them in the context of the available literature and earlier findings. Afterwards in Chapter 3, we give a brief literature overview on Mixed-Integer PDE-constrained optimization and introduce the problem classes, on which our approximation arguments are studied. The second part of the work provides proofs for the approximation steps (a), (b) and (c) and the consequences for optimality (d). We begin with (d) in Chapter 4, introduce a generalized version of the SUR algorithm in Chapter 5 and prove (1.1) for the considered variants of the generalized algorithm and the aforementioned problem classes as well as a signal-processing problem class in Chapters 6, 7 and 9. We make a little excursion in Chapter 8 to present an interesting topological observation, which piqued the author's curiosity. The last part presents computational results to demonstrate the approximation arguments in Chapter 10, summarizes algorithmic considerations in Chapter 11 and provides a concluding discussion in Chapter 12.

1.4 Notation

In $\mathbb{R}^n$, we denote the canonical basis vectors by e_1, …, e_n. Let $\mathcal{X}$ be a Banach lattice with order relation $\leq$, upper bound operation $x \vee y = \sup\{x, y\}$ and lower bound

operation $x \wedge y = \inf\{x, y\}$ for x, $y \in \mathcal{X}$. Then, we denote positive and negative parts for $x \in \mathcal{X}$ by $[x]^+ \coloneqq x \vee 0$ and $[x]^- \coloneqq (-x) \vee 0$.

We use $\mathbb{S}^M \subset \mathbb{R}^M$ to the denote the set containing the extreme points, except zero, of the unit M-simplex, i.e.

$$\mathbb{S}^M \coloneqq \left\{ x \in \mathbb{R}^M : x_i \in \{0, 1\}, \sum_{i=1}^{M} x_i = 1 \right\}.$$

These points are also known as the points satisfying the SOS1 property. We denote the convex hull of a set A by $\operatorname{conv} A$. The characteristic function of a set A is denoted by χ_A. Furthermore, $\mathcal{B}(A)$ denotes the Borel σ-algebra on A. If the considered set A is obvious form the context, we write $\mathcal{B}$. The Lebesgue measure is denoted by λ. For domains Ω and time intervals (t_0, T), we refer to space-time cylinders with the notation $\Omega_T \coloneqq (0, T) \times \Omega$ and $\Omega_{t_0,T} \coloneqq (t_0, T) \times \Omega$. For A being a relatively compact subset of B, we write $A \subset\subset B$.

For a Banach space $\mathcal{X}$, we denote its topological dual by $\mathcal{X}^*$. If not stated otherwise, the duality pairing between $\mathcal{X}$ and $\mathcal{X}^*$ is denoted by $\langle \cdot, \cdot \rangle_{\mathcal{X}^*,\mathcal{X}}$. Similarly, the inner product of a Hilbert space $\mathcal{H}$ is denoted by $\langle \cdot, \cdot \rangle_{\mathcal{H}}$. If a Banach space $\mathcal{X}$ is continuously (compactly) embedded into a Banach space $\mathcal{Y}$, we write $\mathcal{X} \hookrightarrow \mathcal{Y}$ ($\mathcal{X} \hookrightarrow^c \mathcal{Y}$).

If x is the input variable for some function, the operator D_i denotes the partial derivative w.r.t. x_i. The Dirichlet Laplacian is denoted by $-\Delta_D$.

For an optimization problem (P), we denote its feasible set by $\mathcal{F}_{(\mathrm{P})}$.

To avoid cumbersome notation, we denote functions and the corresponding superposition operators with the same symbol and take care that the current use is clear from the context.

For a sequence of $(f^{(n)})_n \subset L^\infty(\Omega_T, \mathbb{R})$, $\liminf f^{(n)}$ abbreviates

$$\liminf_{n\to\infty} \ \sup\{M \in \mathbb{R} : M < f^{(n)}(x) \text{ for a.e. } x \in \Omega_T\}.$$

For $(f^{(n)})_n \subset L^\infty(\Omega_T, \mathbb{R}^n)$ with $n \in \mathbb{N}$, the inequality $M < f^{(n)}(x)$ is understood coordinatewise, i.e. $M < f_i^{(n)}(x)$ for $i \in \{1, \dots, n\}$.

Part I

Background

Chapter 2

Partial outer convexification and Sum-Up Rounding

We summarize the reformulation and relaxations arising from the abstract problem class (MIOCP) in Section 2.1. In Section 2.2, we summarize literature on earlier theoretical findings on the convexification idea and the induced approximation properties in weak topologies. We continue with mixed-integer applications in Section 2.3 and Section 2.4 presents the history on the family of SUR algorithms and their approximation properties in the mixed-integer optimal control context since about 2005.

Before starting, we assign names to $\boldsymbol{\alpha}$ and $\boldsymbol{\omega}$ that highlight their roles in the different relaxations and reformulations of (MIOCP).

Definition 2.1. *Let Ω_T be a bounded domain. We call a function $\boldsymbol{\alpha} \in L^\infty(\Omega_T, \mathbb{R}^M)$ a **relaxed control** if it satisfies $\boldsymbol{\alpha}(s) \in \operatorname{conv} \mathbb{S}^M$ for a.a. $s \in \Omega_T$. We call a function $\boldsymbol{\omega} \in L^\infty(\Omega_T, \mathbb{R}^M)$ a **binary control** if it is a binary-valued relaxed control.*

2.1 Convexification of the abstract problem class

Let us consider the abstract problem class (MIOCP) and derive the relaxations mentioned in Section 1.1. We first convexify the discrete-valued control and then relax the mixed constraint by $\delta \geq 0$ to obtain the equivalent reformulation (case $\delta = 0$) and the feasibility relaxation (case $\delta > 0$):

$$
\begin{aligned}
&\min_{y,u,\boldsymbol{\omega}} \; J(y,u,\boldsymbol{v}) \\
&\text{s.t. } \boldsymbol{\omega}(s) \in \mathbb{S}^M \text{ for a.a. } s \in \Omega_T \\
&\qquad y = S_R(u,\boldsymbol{\omega}),\ u \in \mathcal{U},\ y \in \mathcal{Y} \\
&\qquad -\delta \leq c(y(s),u(s),\boldsymbol{v}(s)) \text{ for a.a. } s \in \Omega_T \\
&\qquad \boldsymbol{v} = \textstyle\sum_{i=1}^M \boldsymbol{\omega}_i v_i.
\end{aligned}
\tag{BC$_\delta$}
$$

The problem class (BC_δ) employs an operator S_R, which is a modification of S that depends on the binary-valued vector $\boldsymbol{\omega}$ instead of the V-valued $\boldsymbol{v}$. For consistency, the state vectors for corresponding vectors $\boldsymbol{v}$ and $\boldsymbol{\omega}$ have to coincide, which is asserted in the assumption below.

Assumption 2.2. *Let $\mathcal{Y}$, $\mathcal{U}$ and V be real Banach spaces. Furthermore,*

1. *let $\mathcal{Y}$, $\mathcal{U}$ be function spaces on a bounded domain $\Omega_T \subset \mathbb{R}^d$ satisfying the continuous embeddings $\mathcal{Y} \hookrightarrow L^1(\Omega_T, Y)$, $\mathcal{U} \hookrightarrow L^1(\Omega_T, U)$ for some Banach spaces Y and U,*
2. *let V be the topological dual of an Asplund space (see Definition B.28),*
3. *let $S(u, \boldsymbol{v}) = S_R(u, \boldsymbol{\omega})$ if $\boldsymbol{v} = \sum_{i=1}^M \boldsymbol{\omega}_i v_i$ for a binary control $\boldsymbol{\omega}$,*
4. *let $c : Y \times U \times V \to \mathbb{R}^{n_c}$ for some $n_c \in \mathbb{N}$ be such that the corresponding superposition operator $c : \mathcal{Y} \times \mathcal{U} \times L^\infty(\Omega_T, V) \to L^\infty(\Omega_T, \mathbb{R}^{n_c})$ is continuous in the first argument.*

Remark 2.3. *Assumption 2.2 2. may seem circumstantial, but by virtue of Theorem B.27 this is what we need to characterize the predual of $L^\infty(\Omega_T, V)$ in the remainder.*

From problem class (BC_δ), the continuous relaxation arises straightforwardly by setting $\delta = 0$ and relaxing the SOS1 property of $\boldsymbol{\omega}$ to convex combinations:

$$\begin{aligned} \min_{y,u,\boldsymbol{\omega}} \; & J(y, u, \boldsymbol{v}) \\ \text{s.t. } & \boldsymbol{\alpha}(s) \in \operatorname{conv} \mathbb{S}^M \text{ for a.a. } s \in \Omega_T \\ & y = S_R(u, \boldsymbol{\alpha}),\ u \in \mathcal{U},\ y \in \mathcal{Y} \\ & 0 \le \boldsymbol{\alpha}_i(s) c(y(s), u(s), v_i) \text{ for a.a. } s \in \Omega_T \text{ for all } i \in \{1, \dots, M\} \\ & \boldsymbol{v} = \sum\nolimits_{i=1}^M \boldsymbol{\alpha}_i v_i. \end{aligned} \tag{RC}$$

Deriving relaxations with the above-described methodology has been introduced in the mixed-integer control community by Sager et al., see e.g. [97, 102]. The notion *partial outer convexification* has been coined in [98]. We summarize earlier work in the subsequent sections and briefly argue for the equivalence and relaxation properties here. The latter arise from the insight that for Banach spaces E, F and any function $f : E \to F$ with superposition operator $f : L^1(\Omega_T, E) \to L^1(\Omega_T, F)$, the identity

$$f(\chi_A g_1 + (1 - \chi_A) g_2) = \chi_A f(g_1) + (1 - \chi_A) f(g_2) \tag{2.1}$$

holds for all $A \in \mathcal{B}$ and g_1, $g_2 \in L^1(\Omega_T, E)$. Before proving equivalences and relaxation properties of the optimization problems above, we define these concepts formally.

Definition 2.4. *Let* (P) *and* (Q) *be optimization problems with corresponding objective functions $J_{(\mathsf{P})}$ and $J_{(\mathsf{Q})}$ as well as corresponding feasible sets $\mathcal{F}_{(\mathsf{P})}$ and $\mathcal{F}_{(\mathsf{Q})}$. Then,*

- *we say that* (P) ***is equivalent to*** (Q) *if there exist surjective mappings* $f : \mathcal{F}_{(\mathsf{P})} \to \mathcal{F}_{(\mathsf{Q})}$ *and* $g : \mathcal{F}_{(\mathsf{Q})} \to \mathcal{F}_{(\mathsf{P})}$ *and constants* $C_1 > 0$, $C_2 \in \mathbb{R}$ *such that* $J_{(\mathsf{Q})}(f(p)) = C_1 J_{(\mathsf{P})}(p) + C_2$ *and* $J_{(\mathsf{Q})}(q) = C_1 J_{(\mathsf{P})}(g(q)) + C_2$ *for all* $p \in \mathcal{F}_{(\mathsf{P})}$ *and all* $q \in \mathcal{F}_{(\mathsf{P})}$,
- *we say that* (P) ***relaxes*** (Q) *or* ***is a relaxation of*** (Q) *if* $\mathcal{F}_{(\mathsf{Q})} \subset \mathcal{F}_{(\mathsf{P})}$ *and* $J_{(\mathsf{Q})} = J_{(\mathsf{P})}$.

It may seem unusual to define equivalence of optimization problems instead of arguing about coinciding solutions of optimization problems. However, as is pointed out in Chapter 4, we cannot expect that the problem class (MIOCP) has any solution in the sense that an integer-valued control function exists that minimizes the objective (locally or globally). We state the connections between the problems introduced above rigorously.

Proposition 2.5. *Let Assumption 2.2 hold. Then, the problem classes* (MIOCP) *and* (BC_0) *are equivalent. Furthermore, the problem classes* (BC_δ) *and* (RC) *relax* (BC_0).

Proof. By combining Assumption 2.2 with the fact that $\boldsymbol{v}(s) \in \operatorname{conv}\{v_1, \dots, v_M\}$ holds for feasible points of (MIOCP) for a.a. $s \in \Omega_T$ and the pointwise a.e. SOS1 property in the first constraint of (BC_0), we obtain the equivalence of (MIOCP) and (BC_0). Obviously, (BC_δ) relaxes (BC_0)for $\delta > 0$. Using the identity (2.1), we obtain the equivalence

$$0 \leq c(y, u, \boldsymbol{v}) \quad \Leftrightarrow \quad 0 \leq \sum_{i=1}^{M} \boldsymbol{\omega}_i c(y, u, v_i)$$

and by the SOS1 property,

$$0 \leq \sum_{i=1}^{M} \boldsymbol{\omega}_i c(y, u, v_i) \quad \Leftrightarrow \quad 0 \leq \boldsymbol{\omega}_i c(y, u, v_i) \text{ for all } i \in \{1, \dots, M\},$$

yielding pointwise almost everywhere equivalence and thus, (RC) relaxes (BC_0). □

Remark 2.6. *The pointwise a.e. constraint* $0 \leq \boldsymbol{\alpha}_i c(y, u, v_i)$ *in* (RC) *yields a so-called* vanishing constraint *in the discretized problem. The class of Mathematical Programs with Vanishing Constraints (MPVCs) has been introduced by Achtziger and Kanzow in* [1] *and studied intensively in the PhD thesis of Hoheisel* [56]. *MPVCs can be reformulated as Mathematical Programs with Complementarity Constraints (MPECs), but exploiting their structure allows for improved constraint qualifications and optimality conditions, see the work by Achtziger, Hoheisel, Izmailov, Kanzow and Solodov* [1, 56–58, 62], *as well as tailored penalty functions, see* [59].

One may wonder why the term *partial outer convexification* contains the word *outer*. This is due to the fact that it was introduced for Ordinary Differential Equation (ODE) systems, where the reformulation of the differential form reads

$$\partial_t y = f(y, u, \boldsymbol{v}) \Leftrightarrow \partial_t y = \sum_{i=1}^{M} \boldsymbol{\omega}_i f(y, u, v_i)$$

and the convexification happens *outside* the function f. Changing the vantage point to the solution operator of the ODE yields the above reformulation for this case.

2.2 The origins of partial outer convexification

The approximation idea behind such convexifications has been prevalent in the optimal control community for a while. The approximation of state-control pairs with convex combinations of feasible state-control pairs or state-control pairs with chattering controls was subject to intensive research some time ago, see e.g. the work by Warga [118, 119], Marchal [83] (*relaxed controls*), Yorke [121] or Cesari [19, Chap. 18.6] (*generalized solutions* and *usual solutions*). A major result is the Filippov–Ważewski theorem [36, 120] below, see Aubin and Cellina's monograph [8, Chap. 2.4 Thm 2] for a proof.

Theorem 2.7 (Filippov–Ważewski theorem)**.** *Let F be a Lipschitz continuous set-valued map into compact subsets of $\mathbb{R}^n$. Then, the solution set of the differential inclusion*

$$\partial_t y(t) \in F(y(t)) \text{ for } t \in [0,T], \quad y(0) = y_0$$

is dense in the solution set of the differential inclusion

$$\partial_t y(t) \in \overline{\operatorname{conv}\{F(y(t))\}}^{\mathbb{R}^n} \text{ for } t \in [0,T], \quad y(0) = y_0$$

if the solution trajectories of the latter differential inclusion are uniformly bounded. ■

Similar to our context, Gamkrelidze noticed that there is a sequence of (continuous) state trajectories, emanating from feasible controls that converges uniformly to a state vector trajectory corresponding to the infimal value of an Optimal Control Problem (OCP) even if no feasible limiting control exists, see [39]. The articles by Frankowska [37] and de Blasi [27] generalize the Filippov–Ważewski theorem to semilinear differential inclusions under similar regularity assumptions to our setting in Section 3.2 on the involved nonlinearity and the assumption that the linear differential operator generates a strongly continuous semigroup. Thus, we can interpret our contributions to the implication (a) in (1.1) as constructive complements of the Filippov–Ważewski theorem that allow to compute the approximants efficiently when having the relaxed solution at hand.

Existence results, which assert the possibility to approximate convex and compact subsets of $(L^\infty(\Omega_T), \sigma(L^\infty, L^1))$ by its extreme points, date further back to the Lyapunov convexity theorem proven first in [78], see also Section 4.1. The proof of a variant of the Lyapunov convexity theorem by Tartar in [114, Thm 3 (2)] is constructive and provides a means to compute the extreme weak* convergent approximants for a given $L^\infty(\Omega_T, \mathbb{R}^M)$-valued function taking values in a convex set $\operatorname{conv} K$. Thus, it offers a possible alternative to the SUR algorithms, which are introduced in Section 2.4.

2.3 Partial outer convexification in mixed-integer applications

From the mid 2000s onwards, Sager and others have developed partial outer convexification to derive reformulations and relaxations of time-dependent MIOCPs, which enabled them to be able to solve a variety of MIOCPs in a computationally efficient way. Regarding applications, we mention the work by Kirches et al. on cruise control of heavy-duty trucks in [67] and time-optimal control of automotive test drives in [69], the work by Logist et al. on multi-objective optimization of automotive test drives and chemical reactor operation and the work by Göttlich et al. on traffic light optimization in [45] and optimal control of transmission lines in [44].

2.4 A brief history of Sum-Up Rounding

The family of SUR algorithms is the rounding technique considered in this work. It serves to bridge the gap between solutions of (RC) and (BC$_\delta$). This section summarizes the development of SUR and the main results in the literature related to it.

What is SUR? The original SUR algorithm, which serves as a building block and starting point for many results in this work, has been introduced by Sager in [97] as *SUR-SOS1* in 2005. A discrete and instructive version is stated and explained below. In Chapter 5, we provide a generalized view on the family of SUR algorithms.

Definition 2.8 (Sum-Up Rounding Algorithm [97, 100])**.** *Let $M \in \mathbb{N}$ and a function $\boldsymbol{\alpha} : \{1, \dots, N\} \to \operatorname{conv} \mathbb{S}^M$ be given. Then, we define the function $\boldsymbol{\omega}(\boldsymbol{\alpha}) : \{1, \dots, N\} \to \{0,1\}^M$, iteratively for $k = 1, \dots, N$ as*

$$\boldsymbol{\omega}_j(k) \coloneqq \begin{cases} 1 & : \quad j = \underset{i \in \{1,\dots,M\}}{\arg\max} \boldsymbol{\alpha}_i(k) + \sum_{\ell=1}^{k-1} \left(\boldsymbol{\alpha}_i(\ell) - \boldsymbol{\omega}_i(\ell) \right), \\ 0 & : \quad \textit{else.} \end{cases}$$

In case of ambiguity, exactly one maximizing index has to be chosen by $\arg\max$*. We often abbreviate $\boldsymbol{\omega} \coloneqq \boldsymbol{\omega}(\boldsymbol{\alpha})$.*

The index $k \in \{1, \dots, N\}$ identifies the current point in time, on which the rounding is performed, and the index $j \in \{1, \dots, M\}$ identifies the discrete value under consideration. First, the entry corresponding to the highest value of $\boldsymbol{\alpha}(1)$ is set to one in $\boldsymbol{\omega}(1)$. The other entries of $\boldsymbol{\omega}(1)$ are set to zero. The algorithm proceeds iteratively and determines the summed-up difference between $\boldsymbol{\alpha}$ and $\boldsymbol{\omega}$, the so-called *integrated control deviation*, until k exclusively plus the value $\boldsymbol{\alpha}(k)$. Then, a maximizing entry of this sum is set to one in $\boldsymbol{\omega}(k)$. Again, the other entries are rounded to zero. Thus, $\boldsymbol{\omega}(k)$ satisfies the SOS1 property for all k. Obviously, the complexity class of the algorithm is $\mathcal{O}(N)$.

When dealing with functions on the domain $[0,T]$, the discrete points in time are replaced by intervals $[t_k, t_{k+1}]$ and the sums by integrals from zero to the interval boundaries to construct a piecewise-constant function $\boldsymbol{\omega} : [0,T] \to \{0,1\}^M$. The key feature of SUR algorithms is that an iterative grid refinement satisfying

$$\max_{k^{(n)}} t^{(n)}_{k^{(n)}+1} - t^{(n)}_{k^{(n)}} \to 0$$

implies

$$\max_{t\in[0,T]} \left\| \int_0^t \boldsymbol{\alpha}(s) - \boldsymbol{\omega}^{(n)}(s)\,\mathrm{d}s \right\|_{\mathbb{R}^M} \to 0.$$

We refer to this property by saying that the control deviation $\boldsymbol{\alpha} - \boldsymbol{\omega}^{(n)}$ is of *vanishing integrality gap*, which is formalized in Section 6.1. Its peculiarity for our analysis is the fact that, under suitable regularity assumptions, it yields norm convergence of corresponding state trajectories from (BC$_\delta$), $(y^{(n)})_n$, to the one of (RC), y, i.e.

$$\|y - y^{(n)}\|_Y \to 0. \tag{2.2}$$

If the objective of an MIOCP is continuous in y and does not depend on the value of the discrete control trajectory, the sequence of corresponding objective values converges to the one of the relaxed problem. These chain of arguments has first been proven by Sager for an ODE setting in [100].

Example We provide a small example for visualization. Consider the functions $g(t) = 0.5 + 0.5\tanh(4t)$ and $h(t) = 0.5$. Both are bounded functions and can be rewritten as time-varying convex combinations of the constant functions $v_1 \equiv 1$ and $v_2 \equiv 0$. We have computed convex coefficients $\boldsymbol{\alpha}$ for them and applied the SUR algorithm to interval-wise averaged versions of them for two grids ($N = 32$ and $N = 128$). The functions and their SUR approximants are plotted in Fig. 2.1, the right column of which is published in [80, Fig. 1]. We observe that the roundings cannot give any reasonable approximation in the norm topology of any L^p-space.

We consider simple Initial Value Problems (IVPs) for g and h: $y_g(-5) = 0, \partial_t y_g(t) = g(t)$ and $y_h(-5) = 0, \partial_t y_h(t) = h(t)$. In Fig. 2.2, we observe that, as claimed in (2.2), the state trajectories y_g and y_h are approximated in norm by the state trajectories resulting from the SUR approximations of the right hand sides, which we call y_g^{SUR} and y_h^{SUR}.

Extensions Gerdts and Sager have investigated MIOCPs, in which the right hand sides of the ODE and an additional constraint, comprising a Differential-Algebraic Equation (DAE) of index one, depend on discrete controls and are convexified, see [43]. They apply the implicit function theorem under suitable regularity assumptions on the algebraic equation, which allows them to transfer the approximation properties to the DAE setting. We note that the *variable time transformation* described in [42, 43] may be interpreted

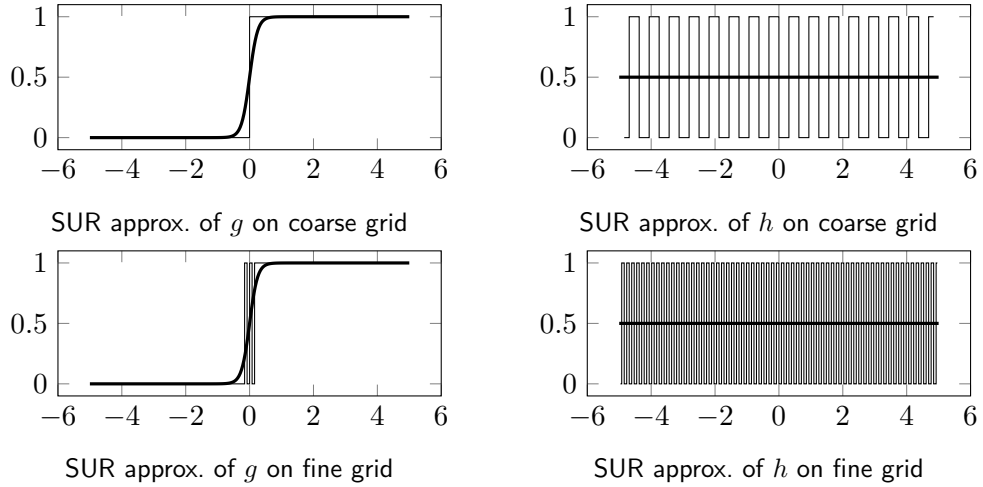


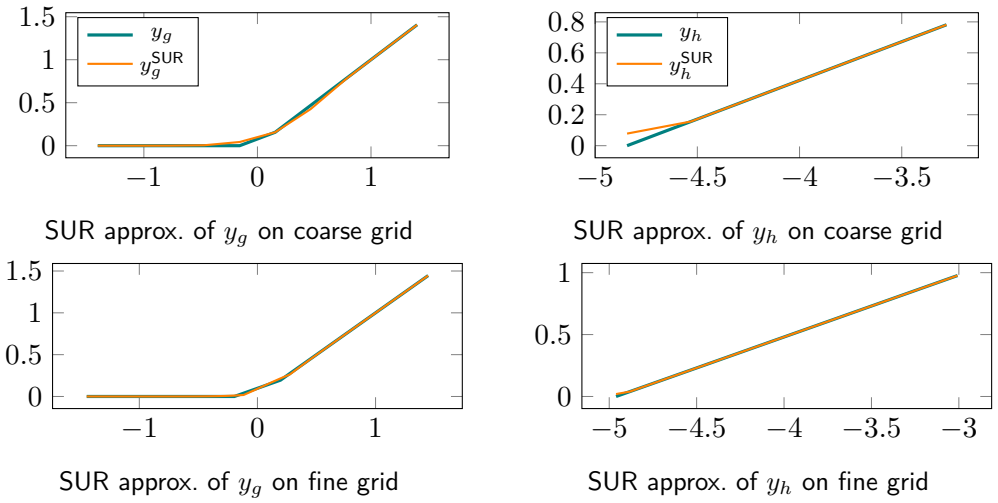
Figure 2.1: SUR approximation (thin) for functions (thick) g (left) and h (right) on coarse (top) and fine grid (bottom). The plots in the right column have been published in [80, Fig. 1].

as an instance of Tartar's construction methodology mentioned in Section 2.2 in the one-dimensional case.

Kirches [66] combined the partial outer convexification and SUR with the *real-time iteration scheme* by Diehl [30] to solve *Mixed-Integer Model Predictive Control* problems.

To include path constraints that also depend on the discrete controls into the problem (MIOCP), one has to avoid the problem that a rounded control may lead to arbitrarily large constraint violations. Ensuring that a sequence of control deviations is of vanishing integrality gap is not sufficient to drive the infeasibility to zero in an L^∞ sense. In [68, 71, 82] by Lenders, Kirches and the author, a modification of the SUR algorithm is introduced that is able to drive the infeasibility to zero. We analyze this variant in Sections 6.4 and 6.5. Recent work by Bock et al. [13] uses this result when studying MPVCs arising from the conversion of implicitly switched systems into explicitly switched ones. Further types of constraints have been investigated by Sager in [98].

An alternative rounding algorithm, Next-Forced Rounding (NFR), is introduced and investigated by Jung in [64]. Its rounding strategy abides the bound $\max_{t\in(0,T)} \| \int_0^t (\boldsymbol{\alpha} - \boldsymbol{\omega})\|_\infty \leq h$ with h denoting the mesh size of the rounding grid, see [64, Prop. 4.8]. This bound is superior to the ones that can be achieved for SUR algorithms, which is $\mathcal{O}(\log(M))h$ for the original variant of the algorithm, see [68] and the proof in Section 6.3. NFR enforces rounding of critical entries to hold this bound. However, its computational complexity is in $\mathcal{O}(N^2)$ where N denotes the number discretization cells, see [64, Rem. 4.13]. Furthermore, NFR cannot be modified as easily as SUR to treat additional mixed state-control constraints as we carve out in Section 5.4.



SUR approx. of y_g on coarse grid

SUR approx. of y_h on coarse grid

SUR approx. of y_g on fine grid

SUR approx. of y_h on fine grid

Figure 2.2: SUR approximation of trajectories y_g (left) and y_h (right) on coarse (top) and fine grid (bottom).

Generalizations As already mentioned in Section 1.1, the convexification and application of a SUR algorithm can be regarded as a special case of the general methodology to split the process to solve a MIOCP or, more precisely, to find an approximate solution of (BC_δ) into the following two parts:

1. solve a relaxation, (RC) in the case of partial outer convexification, of the underlying MIOCP;

2. solve a rounding problem to obtain a corresponding integer control, feasible for (BC_δ) in the case of partial outer convexification.

The approximation of the solution of (RC) in terms of feasibility and optimality are implied by a vanishing integrality gap on refined rounding grids. As pointed out by Sager et al. in [65, 101, 123], the rounding problem, also named Combinatorial Integral Approximation (CIA) in the aforementioned references, can be reformulated as a Mixed-Integer Linear Program (MILP) and different rounding strategies and estimates can be obtained by modifying the MILP. Furthermore, binary controls can be synthesized from existing ones obtained with different rounding strategies by means of recombination heuristics, see [123]. Such heuristics are of course important if the mesh size of the rounding grid is fixed or confined to narrow limits.

SUR for Mixed-Integer PDE-Constrained Optimization Problems (MIPDECOs) In their article [51], Hante and Sager have transferred SUR and the approximation property from Sager [97, 100] to time-dependent Partial Differential Equations (PDEs) with linear differential operators generating strongly continuous semigroups. They present an

algorithm to compute integer control trajectories from the relaxed ones on an iteratively refined grid and show the desired convergence of the state trajectories as introduced in (2.2) under the assumption that the control deviation is of vanishing integrality gap. As the regularity assumptions on the solution trajectory and the behavior of the nonlinear term from [50] are hard to satisfy and check for hyperbolic PDEs, Hante has shown the results to hold also under weaker regularity assumptions, which are more suitable for hyperbolic PDEs but still require some degree of differentiability of the involved quantities, see [49, 50].

In Section 9.1, we prove that having a sequence of vanishing integrality gap yields norm convergence of the corresponding sequence of state vectors under weaker regularity assumptions. In particular, we relax the assumptions on the nonlinear part of the PDE evaluated at the relaxed solution. This result has been published by Kirches and the author in [80].

In their article [48], Hahn and Sager employ ideas from SUR to approximation problems of the form

$$\min_{\boldsymbol{v}} \|y(u^*, \boldsymbol{v}^*) - y(u^*, \boldsymbol{v})\|_Y$$

where $(u, \boldsymbol{v}) \mapsto y(u, \boldsymbol{v})$ is assumed to be the solution of some underlying PDE with respect to continuous- and discrete-valued control inputs u and $\boldsymbol{v}$ and $(u^*, \boldsymbol{v}^*)$ is the known solution of a relaxed problem. Theorem 5.4 in [48] shows that if the mapping $(u, \boldsymbol{v}) \mapsto y(u, \boldsymbol{v})$ is uniformly locally Lipschitz continuous in the second argument, one can bind the integer approximation error everywhere with the sum of the approximation error on a finite grid and an $\mathcal{O}(d)$-term with d denoting the maximum interior distance between two grid points. Furthermore, [48] states a heuristic algorithm called *Simple Pivot Search* to approximate binary controls from relaxed continuous ones.

Having employed the SUR algorithm for time-dependent PDEs, it is natural to ask if the algorithm can be generalized to multi-dimensional domains or space-time cylinders. We prepare a multi-dimensional generalization in our formulation of the SUR algorithm in Chapter 5 and answer the question positively in Section 7.2 by showing that the dimension-dependent implication (b) in (1.1) can be proven for multi-dimensional domains under mild assumptions on the employed grid refinement strategy.

Chapter 3

Mixed-Integer PDE-Constrained Optimization

Most of the MIOCPs considered in this work are instances of the prototypical class (MIPDECO) below.

$$\begin{aligned} \min_{y,u,\boldsymbol{v}} \; & J(y,u) \\ \text{s.t.} \quad & E(y) = f(y,u,\boldsymbol{v}) \\ & 0 \leq c(y(t,x),u(t,x),\boldsymbol{v}(t,x)) \quad \text{a.e.} \\ & \boldsymbol{v}(t,x) \in \{v_1,\dots,v_M\} \quad \text{a.e.} \end{aligned} \tag{MIPDECO}$$

Of course, (MIPDECO) is underspecified so far, but it reveals the basic structure of the state equation we assume. The state equation, usually a PDE or ODE and its initial and boundary values, is restricted as follows: the differential operators act on the state vector only and are contained in the term $E(y)$. Furthermore, we assume a nonlinear right hand side f that is good-natured, e.g. Lipschitz continuous in y. Additionally, (MIPDECO) features a mixed constraint, which is assumed to be continuous in terms of the state vector. The forthcoming sections make this setup rigorous for two classes of semilinear evolution equations and a class of elliptic equations.

In several talks, Leyffer has classified MIPDECOs into *mesh-dependent* and *mesh-independent* problems.[1,2] With respect to this taxonomy, this work investigates mesh-dependent or distributed problems, i.e. problems where the number of discrete-valued variables depends on the mesh size of the discretization of some infinite-dimensional process or relationship. This is not the case for mesh-independent MIPDECOs where the number of discrete-valued variables does not change when the mesh that discretizes the infinite-dimensional process is refined. In principle, mesh-independent problems can be approached by algorithms from the MINLP community. We emphasize that MINLPs

[1] `https://wiki.mcs.anl.gov/leyffer/images/8/8b/SvenISMP15.pdf`, visited on July 17, 2019

[2] `http://coral.ise.lehigh.edu/usmex2016//files/2016/04/talks/leyffer.pdf`, visted July 17, 2019

are generally very challenging and one should refrain from belittling mesh-independent MIPDECOs as the *easier cases*. On the contrary, the theory presented in this work shows that we can sometimes gain a computationally efficient algorithmic framework to approach mesh-dependent problems by analyzing the considered problem class from the infinite-dimensional point of view although this may seem paradoxical at the beginning.

We continue this chapter with a brief literature overview on MIPDECOs that focuses mainly on mesh-dependent MIPDECOs and continue with three sections that introduce precise setups for the problem classes that are subject to our analysis in the remainder.

3.1 Literature overview

Although not in the focus of this work, we mention two examples of mesh-independent problems that have been addressed in the literature. In [9], Bachmann et al. consider a controlled heat equation, in which the discrete controls model different insulation materials for the boundaries of the domain. The discrete optimization is solved with a branch-and-bound method that is accelerated by using a POD-based reduced model, see [96], to solve the parabolic IVP. The second one is presented by Buchheim et al. in [17] and constitutes an interplay of PDE-theory and a discrete optimization algorithm. The authors employ an *Outer Approximation* algorithm to solve a class of OCPs, which are constrained by semilinear elliptic equations with a finite number of pure binary control inputs, within finitely many iterations by means of a suitable cutting-plane generation strategy. The cutting-plane generation strategy is based on a projection of the problem to the control space and employing pointwise concavity and submodularity of the evaluation of the *control-to-state operator*. The results raise the expectation that a large number of binary controls can be handled for the considered problem class because it evades the drawback of the aforementioned SOS1-based reformulations that the number of feasible integer states, M in this work, may grow exponentially with the number of possible discrete events.

Probably the most prominent class of problems that can be cast as mesh-dependent MIPDECOs is topology optimization. We name the monographs of Bendsøe and Sigmund [12] and Pironneau [90] for extensive information on topology and shape optimization. Roughly, topology optimization approaches can be categorized into density-based and shape optimization-based approaches, see [106]. Shape optimization, see e.g. [54, 61], often circumvents MIPDECOs formulations and employs the shape derivative information to modify the characteristic functions, i.e. the assignment of the parts of the mesh (nodes, patches) belonging to the different regions. Density-based approaches often solve a relaxation of the MINLP arising from discretization and use penalty terms in the objective to obtain a solution that is already *almost* binary-valued. We refer to [106] for a comparative view on different approaches to topology optimization. Another example of a work on distributed topology optimization where a distributed integer formulation is used is given in [105]. The PDE-constrained problem with binary variables is

relaxed to continuous ones and an exact rounding is performed by means of a level set function whose zero set defines the separating interface implicitly. Then, state, adjoint and the shape gradient are computed for the current rounding and the level set function is updated in the direction of the shape gradient by means of an advection-diffusion process over a fixed time horizon. Furthermore, the article [48], see also Section 2.4, compares different formulations to approach a discretized source inversion and a discretized topology optimization problem with different problem formulations.

Regarding the treatment of discrete-valued distributed controls, we mention the work of Clason and Kunisch who use an L^0-type control regularizer to promote that a continuously-valued control takes discrete values for elliptic control problems in [22], i.e. their method falls into the aforementioned category of density-based approaches. They prove a *generalized multi-bang* principle that yields conditions under which it is possible to characterize the parts of the domain where the control takes the desired discrete values, the so-called *bangs*, see [22, Prop. 2.3]. The approach is extended to state equations, in which the control enters as a diffusion coefficient in the differential operator or a bilinear term in the state equation in [23], vector-valued *bangs* in [24] and total variation regularization in [21].

Another prevalent class of mesh-dependent MIPDECOs are network flow problems where the flow on an edge is determined by a hyperbolic PDE as e.g. in optimal gas transport problems. We mention the report [38], which summarizes work on network flow optimization problems from different application areas with a focus on the question how to select an abstraction level for surrogate models and optimization techniques to master the computational demand. In this context, the PDE is often fully discretized or replaced by a surrogate model and the discretized problem is optimized using tailored integer programming techniques, see e.g. [70, 84, 89, 112] and the references therein. We also mention Gugat et al.'s work [47], which shows good results with the suboptimal *instantaneous control approach* using a surrogate ODE and successive solutions of linear Integer Programs (IPs) and the work by Hante [49, 50], which employs relaxation-based techniques as discussed in this work for control of a hyperbolic system.

Although dealing with a signal reconstruction problem, in which no differential equation is involved, we mention the work of Buchheim et al. [16]. They propose a *Branch-and-Bound* algorithm to minimize a quadratic signal reconstruction objective, in which a discrete-valued activation function is convolved with a filter function to reconstruct a given target signal. The resulting discrete control trajectories exhibit a chattering behavior that resembles the one resulting from solving (RC) and applying SUR algorithms. Furthermore, the option to switch at very high frequencies is a characteristic in signal processing, which allows the implementation of such trajectories even for fine discretizations. Thus, we also investigate our relaxation-based methodology for the problem posed in [16] in Section 10.2. Chapter 4 shows that it is not incidental that the proposed methodology is not restricted to systems governed by differential equations.

3.2 A problem class with distributed integer controls in time

We combine results of Kirches, Lenders and the author [68, 82] for the ODE case, i.e. $E(y) \coloneqq \partial_t y$ and $y(0) = y_0$, including mixed constraints depending on the discrete-valued variable with results by Kirches and the author [80] on semilinear IVPs, i.e. $E(y) = \partial_t y + Ay$ where the operator A generates a strongly continuous semigroup. Thus, we provide a class of MIOCPs that covers both settings. For a brief statement of the employed results from semigroup theory, we refer to Appendix B.2.

$$\begin{aligned} \min_{y,u,\boldsymbol{v}} \; & J(y,u,\boldsymbol{v}) \\ \text{s.t.} \; & \partial_t y + Ay = -f(y,u,\boldsymbol{v}) \\ & y(0) = y_0 \\ & \boldsymbol{v}(t) \in \{v_1,\dots,v_M\} \qquad \text{a.e. in } [0,T] \\ & 0 \le c(y(t),u(t),\boldsymbol{v}(t)) \quad \text{a.e. in } [0,T] \end{aligned} \tag{MIPEVO-T}$$

To state a functional analytic setting that covers the IVPs in (MIPEVO-T) as well as its reformulation and relaxations, we prepone the convexified formulation of the IVP:

$$\begin{aligned} \partial_t y + Ay &= -\sum_{i=1}^{M} \boldsymbol{\alpha}_i f(y,u,v_i) \\ y(0) &= y_0. \end{aligned} \tag{3.1}$$

We introduce the functional analytic setting to discuss (MIPEVO-T).

Assumption 3.1 (Setting of (MIPEVO-T))**.**

1. *Let Y be a real Banach space and $A : D(A) \to Y$ be a linear operator that generates a strongly continuous semigroup on Y. Let $y_0 \in Y$.*
2. *Let U be a Banach space and u be restricted to $L^2((0,T),U)$, i.e. $u \in L^2((0,T),U)$.*
3. *Let V be the topological dual of an Asplund space. Let $\{v_1,\dots,v_M\} \subset V$ and $\boldsymbol{v}$ be restricted to $L^\infty((0,T),V)$, i.e. $\boldsymbol{v} \in L^\infty((0,T),V)$.*
4. *Let $f : Y \times U \times \{v_1,\dots,v_M\} \to Y$ be uniformly Lipschitz continuous in the first argument and jointly continuous in the first and second argument. Furthermore, let $f(0,u,v_i) \in L^1((0,T),Y)$ for all $u \in L^2((0,T),U)$ and all $i \in \{1,\dots,M\}$.*

Proposition 3.2. *Let Assumption 3.1 hold. Then,* (3.1) *admits a unique mild solution $y \in C([0,T],Y) =: \mathcal{Y}$ for all $u \in L^2((0,T),U)$, all relaxed controls $\boldsymbol{\alpha} \in L^\infty((0,T),\mathbb{R}^M)$ and all $y_0 \in Y$.*

Proof. We defer the proof to Proposition B.16. □

Remark 3.3. *For the ODE setting, we have $Y = \mathbb{R}^n$ and a typical choice for a PDE setting would be $Y = L^2(\Omega)$ for some bounded domain Ω with smooth or Lipschitz boundary. In the latter case, A may be unbounded and $D(A)$ denotes its domain.*

Remark 3.4. *An explicit dependence of the nonlinearity f on $[0,T]$ would be possible as well, see* [80]*, but has been omitted as this only bloats the PDE analysis and plays a minor role w.r.t. our main chain of approximation arguments* (1.1)*.*

We have deliberately chosen a restrictive setting for the nonlinear right hand side f in the IVP in Assumption 3.1, which allows us to treat the PDE as an abstract ODE by means of semigroup theory. The advantage of this setting is that it includes several hyperbolic PDEs as well. If one wants to consider more difficult nonlinearities, it makes sense to restrict oneself to differential operators that exhibit maximal parabolic regularity, which is what we do in Section 3.4.

The partial outer convexification reformulation ($\mathsf{BC}_0^{\text{EVO-T}}$) of (MIPEVO-T) and the corresponding feasibility relaxation ($\mathsf{BC}_\delta^{\text{EVO-T}}$) are stated below:

$$
\begin{aligned}
\min_{y,u,\boldsymbol{\omega}} \; & J(y,u,\boldsymbol{v}) \\
\text{s.t. } & \partial_t y + Ay = -\textstyle\sum_{i=1}^M \boldsymbol{\omega}_i f(y,u,v_i) \\
& y(0) = y_0 \\
& -\delta \le \boldsymbol{\omega}_i(t) c(y(t),u(t),v_i) \quad \text{a.e. in } [0,T] \text{ for } 1 \le i \le M \\
& \boldsymbol{v} = \textstyle\sum_{i=1}^M \boldsymbol{\omega}_i v_i \\
& \boldsymbol{\omega}(t) \in \mathbb{S}^M \qquad\qquad \text{a.e. in } [0,T].
\end{aligned}
\tag{$\mathsf{BC}_\delta^{\text{EVO-T}}$}
$$

The following proposition, proven in Lenders' and Sager's PhD theses [71, 97], states the equivalence of the reformulation and provides the reconstruction formula for solutions of (MIPEVO-T) from solutions of ($\mathsf{BC}_0^{\text{EVO-T}}$). It follows from Proposition 2.5.

Proposition 3.5. *Assume that Assumption 3.1 holds. Then,* (MIPEVO-T) *is equivalent to* ($\mathsf{BC}_0^{\text{EVO-T}}$) *with the correspondence mapping*

$$
\mathcal{F}_{(\mathsf{BC}_0^{\text{EVO-T}})} \ni (y,u,\boldsymbol{\omega}) \mapsto \left(y, u, \sum_{i=1}^M \boldsymbol{\omega}_i v_i \right) \in \mathcal{F}_{(\text{MIPEVO-T})}.
$$

Proof. Once Assumption 2.2 is satisfied, the claim follows from the surjectivity of the correspondence mapping in Proposition 2.5. The surjectivity follows from Carathéodory's theorem, see Theorem B.33. Assumption 2.2 follows from the identity (2.1) and the equivalence of the IVPs under the correspondence mapping, which can be observed from the definitions, see also [71, Prop. 6.6]. □

As demonstrated in Section 2.1 and conducted by Kirches, Lenders and the author in [68, 71, 82], the continuous relaxation ($\mathsf{RC}^{\text{EVO-T}}$) arises by weakening the SOS1 property of the binary coefficients $\boldsymbol{\omega}$ to convex combinations. We denote the coefficients of the convex combinations by $\boldsymbol{\alpha}$ and arrive at the following OCP, of which different variants

were studied by Sager and Hante in [50,51] and Kirches and the author in [68,71,80,82]:

$$\begin{aligned}
\min_{y,u,\boldsymbol{\alpha}}\ & J(y,u,\boldsymbol{v}) \\
\text{s.t. } & \partial_t y + Ay = \textstyle\sum_{i=1}^{M} -\boldsymbol{\alpha}_i f(y,u,v_i) \\
& y(0) = y_0 \\
& 0 \leq \boldsymbol{\alpha}_i(t) c(y(t),u(t),v_i) \text{ a.e. in } [0,T] \text{ for } 1 \leq i \leq M \\
& \boldsymbol{v} = \textstyle\sum_{i=1}^{M} \boldsymbol{\alpha}_i v_i \\
& \boldsymbol{\alpha}(t) \in \operatorname{conv} \mathbb{S}^M \qquad \text{a.e. in } [0,T].
\end{aligned} \tag{RC$^{\text{EVO-T}}$}$$

To complete our introduction of the problem class, we consider the IVPs that constrain (MIPEVO-T) and (BC$_\delta^{\text{EVO-T}}$) for $\delta \geq 0$ and provide a short lemma to assert that their solution theory is covered by Assumption 3.1, too. The IVPs read

$$\begin{aligned}
\partial_t y + Ay &= -f(y,u,\boldsymbol{v}) \\
y(0) &= y_0
\end{aligned} \tag{3.2}$$

for (MIPEVO-T) and

$$\begin{aligned}
\partial_t y + Ay &= -\sum_{i=1}^{M} \boldsymbol{\omega}_i f(y,u,v_i) \\
y(0) &= y_0
\end{aligned} \tag{3.3}$$

for (BC$_\delta^{\text{EVO-T}}$) where $\boldsymbol{\omega} \in L^\infty((0,T),\mathbb{R}^M)$ is a binary control.

Lemma 3.6. *Let Assumption 3.1 hold. Then, the IVP*

1. (3.3) *admits a unique mild solution for all $u \in L^2((0,T),U)$, all binary controls $\boldsymbol{\omega} \in L^\infty((0,T),\mathbb{R}^M)$ and all $y_0 \in Y$.*

2. (3.2) *admits a unique mild solution for all $u \in L^2((0,T),U)$, all $\{v_1,\dots,v_M\}$-valued $\boldsymbol{v} \in L^\infty((0,T),V)$ and all $y_0 \in Y$.*

Proof. The first claim holds true as the admissible $\boldsymbol{\omega}$ are a subset of the admissible $\boldsymbol{\alpha}$ in Assumption 3.1. To show the second claim, we notice that the superposition operator of the function f maps to $L^1((0,T),Y)$ and employ the identity (2.1) to obtain that the IVPs (3.2) and (3.3) are equivalent with the identification from Proposition 3.5. □

3.3 A problem class with distributed integer controls in space

(MIPEVO-T) is *the* problem class for which partial outer convexification and the effect of SUR algorithms have been investigated in the literature. The adaption of the methodology for the problem class (MIPELL) below is a recent development put forward by the author in [79]. Therein, a class of MIOCPs constrained by elliptic Boundary Value

Problems (BVPs) with homogeneous Dirichlet boundaries has been investigated in the absence of mixed constraints. We add the mixed constraints and arrive at the MIOCP

$$\begin{aligned} \min_{y,u,\boldsymbol{v}} \; & J(y,u,\boldsymbol{v}) \\ \text{s.t.} \quad & Ay = f(u,\boldsymbol{v}) \\ & \boldsymbol{v}(x) \in \{v_1,\dots,v_M\} \qquad \text{a.e. in } \Omega \\ & 0 \le c(y(x),u(x),\boldsymbol{v}(x)) \text{ a.e. in } \Omega. \end{aligned} \tag{MIPELL}$$

The functional analytic setting to discuss (MIPELL) is given in the following assumption.

Assumption 3.7 (Setting of (MIPELL))**.**

1. *Let Ω be a bounded domain, let $\mathcal{V}$ be a Hilbert space on Ω. Let the so-called Gelfand triple $\mathcal{V} \hookrightarrow^c L^2(\Omega) \cong L^2(\Omega)^* \hookrightarrow^c \mathcal{V}^*$ hold with continuous, dense and compact embeddings.*

2. *Let $\mathcal{U} \hookrightarrow L^2(\Omega,U)$ and U be Banach spaces and u be restricted to $\mathcal{U}$, i.e. $u \in \mathcal{U}$.*

3. *Let $A : \mathcal{V} \to \mathcal{V}^*$ be a linear mapping with bounded inverse. In particular, there exists $C > 0$ such that for all $f \in \mathcal{V}^*$ the state equation*

$$Ay = f$$

admits a unique solution $y \in \mathcal{V}$ for which the estimate

$$\|y\|_{\mathcal{V}} \le C\|f\|_{\mathcal{V}^*}$$

holds.

4. *Let V be the topological dual of an Asplund space. Let $\{v_1,\dots,v_M\} \subset V$ and $\boldsymbol{v}$ be restricted to $L^\infty(\Omega,V)$, i.e. $\boldsymbol{v} \in L^\infty(\Omega,V)$.*

5. *Let $f : U \times V \to \mathbb{R}$ be such that the superposition operators*

$$f(u,v_i)(x) \coloneqq f(u(x),v_i)$$

are continuous mappings $f(\cdot,v_i) : \mathcal{U} \to L^2(\Omega)$ for all $i \in \{1,\dots,M\}$.

We provide a well-known setting that satisfies the regularity in Assumption 3.7.

Proposition 3.8. *Let Ω be a bounded domain. Let $\mathcal{V} = H_0^1(\Omega)$. Let $A = -\Delta_D$. Then, the claims 1. and 3. in Assumption 3.7 are satisfied.*

Proof. The embeddings and their properties in Assumption 3.7 follow from the Sobolev embedding $H^1(\Omega) \hookrightarrow^c L^2(\Omega)$, see e.g. [94, Thm 7.29]. Existence and uniqueness follow from the variational formulation and the Lax–Milgram lemma (see Theorem B.6), i.e. we work with weak solutions here. We refer to [94, Chap. 9.2] for an instructive derivation. □

The partial outer convexification reformulation ($\mathsf{BC}_0^{\mathsf{ELL}}$) of (MIPELL) and the corresponding feasibility relaxation ($\mathsf{BC}_\delta^{\mathsf{ELL}}$) are stated below.

$$\begin{aligned}
\min_{y,u,\boldsymbol{\omega}}\ & J(y,u,\boldsymbol{v}) \\
\text{s.t.}\quad & Ay = \textstyle\sum_{i=1}^M \boldsymbol{\omega}_i f(u,v_i) \\
& -\delta \le \boldsymbol{\omega}_i(x) c(y(x),u(x),v_i) \ \text{a.e. in } \Omega \text{ for } 1 \le i \le M \\
& \boldsymbol{v} = \textstyle\sum_{i=1}^M \boldsymbol{\omega}_i v_i \\
& \boldsymbol{\omega}(x) \in \mathbb{S}^M \qquad\qquad \text{a.e. in } \Omega
\end{aligned} \tag{$\mathsf{BC}_\delta^{\mathsf{ELL}}$}$$

As for (MIPEVO-T), the equivalence of (MIPELL) to ($\mathsf{BC}_0^{\mathsf{ELL}}$) is a special case of Proposition 2.5, which we state below.

Proposition 3.9. *Assume that Assumption 3.7 holds. Then,* (MIPELL) *is equivalent to* ($\mathsf{BC}_0^{\mathsf{ELL}}$) *with the correspondence mapping*

$$\mathcal{F}_{(\mathsf{BC}_0^{\mathsf{ELL}})} \ni (y,u,\boldsymbol{\omega}) \mapsto \left(y,u,\sum_{i=1}^M \boldsymbol{\omega}_i v_i\right) \in \mathcal{F}_{(\mathsf{MIPELL})}.$$

Proof. Once Assumption 2.2 is satisfied, the claim follows from the surjectivity of the correspondence mapping Proposition 2.5. The surjectivity follows from Carathéodory's theorem (see Theorem B.33). Assumption 2.2 follows from the identity (2.1) and the choices $S(u,\boldsymbol{v}) \coloneqq A^{-1} f(u,\boldsymbol{v})$ and $S_R(u,\boldsymbol{\omega}) \coloneqq A^{-1} \sum_{i=1}^M \boldsymbol{\omega}_i f(u,v_i)$. □

Again, we derive the continuous relaxation by weakening the SOS1 property of $\boldsymbol{\omega}$ to convex combinations, denote the new coefficients by $\boldsymbol{\alpha}$ and obtain the following OCP, which was studied by Kirches and the author in [79] in absence of the mixed constraint.

$$\begin{aligned}
\min_{y,u,\boldsymbol{\alpha}}\ & J(y,u,\boldsymbol{v}) \\
\text{s.t.}\quad & Ay = \textstyle\sum_{i=1}^M \boldsymbol{\omega}_i f(u,v_i) \\
& -\delta \le \boldsymbol{\alpha}_i(x) c(y(x),u(x),v_i) \ \text{a.e. in } \Omega \text{ for } 1 \le i \le M \\
& \boldsymbol{v} = \textstyle\sum_{i=1}^M \boldsymbol{\alpha}_i v_i \\
& \boldsymbol{\alpha}(x) \in \operatorname{conv} \mathbb{S}^M \qquad\qquad \text{a.e. in } \Omega
\end{aligned} \tag{$\mathsf{RC}^{\mathsf{ELL}}$}$$

3.4 A problem class with distributed integer controls in time and space

In this section, we introduce a problem class with discrete-valued controls that are distributed in time and space. To the best of the author's knowledge, this has not been

considered in the literature so far. The problem class reads

$$\begin{aligned} \min_{y,u,\boldsymbol{v}} \; & J(y,u,\boldsymbol{v}) \\ \text{s.t.} \;\; & \partial_t y + Ay = -f(y,u,\boldsymbol{v}) \\ & y((0,\cdot)) = y_0 \\ & y|_{\partial\Omega} = 0 \\ & \boldsymbol{v}(s) \in \{v_1,\dots,v_M\} \qquad \text{a.e. in } \Omega_T \\ & 0 \le c(y(s),u(s),\boldsymbol{v}(s)) \;\; \text{a.e. in } \Omega_T. \end{aligned} \tag{MIPEVO-TX}$$

Here, we assume the structure $f(y,u,\boldsymbol{v}) = f^a(y,\boldsymbol{v}) + f^b(y,u)$ where the functions f^a and f^b, and as a consequence also f, are superposition operators induced by real-valued functions. The detailed assumptions for the problem class (MIPEVO-TX), in particular the IVP, are provided below. They resemble the setting considered by Raymond and Zidani in [93]. In contrast to [93], we have simplified the setting e.g. by replacing the Robin boundary conditions by homogeneous Dirichlet boundary conditions. However, we do not expect any major changes in our arguments if one considers the original setting and the corresponding assumptions from [93]. In the following, we impose three assumptions on the problem class. The first two assumptions serve to provide a solution theory for the IVP that constraints the problem class (MIPEVO-TX) in the space

$$\mathcal{W} \coloneqq W((0,T)) = \left\{ y \in L^2((0,T),\mathcal{V}) : \partial_t y \in L^2((0,T),\mathcal{V}^*) \right\} \hookrightarrow^c L^2((0,T),\mathcal{H}) =: \mathcal{Y},$$

where the compact embedding follows from the Aubin-Lions-Simon lemma. The last assumption is only required to show implication (c) in (1.1) later. As in Section 3.2, we state a functional analytic setting that covers the IVPs in (MIPEVO-TX) as well as its reformulation and relaxations and prepone the convexified reformulation:

$$\begin{aligned} \partial_t y + Ay &= -\sum_{i=1}^{M} \boldsymbol{\alpha}_i f^a(y,v_i) - f^b(y,u) \\ y(0) &= y_0 \\ y|_{\partial\Omega} &= 0. \end{aligned} \tag{3.4}$$

Assumption 3.10 (Domain and differential operator, leaning on [93, Sect. 2])**.**

1. *Let $d \ge 2$ and let $\Omega \subset \mathbb{R}^d$ be a bounded domain.*
2. *Let $q > \frac{d}{2} + 1$ and u be restricted to $L^q(\Omega_T)$, i.e. $u \in L^q(\Omega_T)$.*
3. *Let $\mathcal{V} \hookrightarrow^c \mathcal{H} \cong \mathcal{H}^* \hookrightarrow^c \mathcal{V}^*$ with the choices $\mathcal{V} \coloneqq H_0^1(\Omega)$, $\mathcal{H} \coloneqq L^2(\Omega)$ and $\mathcal{V}^* \coloneqq H^{-1}(\Omega)$.*
4. *Let A be a second-order differential operator defined by the bilinear form*

$$\langle Ay, \varphi \rangle_{\mathcal{V}^*,\mathcal{V}} \coloneqq \sum_{i=1}^{d} \sum_{j=1}^{d} \int_\Omega a_{ij}(x) D_j y(x) D_i \varphi(x) \,\mathrm{d}x$$

for $y, \varphi \in \mathcal{V}$. We assume that the $\mathbb{R}^{d\times d}$-valued coefficient function a satisfies $a_{ij} \in C(\bar{\Omega})$ for all $i, j \in \{1, \dots, N\}$ and the following symmetry and boundedness conditions

$$\begin{aligned} a_{ij}(x) &= a_{ji}(x) \text{ for all } i, j \in \{1, \dots, N\}, \\ m_0\|\xi\|^2 \leq \xi^T a(x)\xi &\leq M_0\|\xi\|^2. \end{aligned}$$

for all $x \in \bar{\Omega}$ and all $\xi \in \mathbb{R}^d$ and some $0 < m_0 \leq M_0$. [3]

Assumption 3.11 (Nonlinearity setup, leaning on Ass. (A1) in [93, Sect. 2.2])**.**

1. *Let $u \in \mathbb{R} \times \mathbb{R}^M \cong \mathbb{R}^{1+M}$ and $f : \mathbb{R} \times \mathbb{R} \times \{v_1, \dots, v_M\} \to \mathbb{R}$ be defined as*

$$f(y, u, v_i) \coloneqq f^a(y, v_i) + f^b(y, u)$$

for $i \in \{1, \dots, M\}$.

2. *Let $f^a(\cdot, v_i) \in C(\mathbb{R})$ for all $i \in \{1, \dots, M\}$ and $f^b \in C(\mathbb{R}^2)$. Let $f^a(\cdot, v_i) \in C^1(\mathbb{R})$ for all $i \in \{1, \dots, M\}$ and $f^b(\cdot, u) \in C^1(\mathbb{R})$ for all $u \in \mathbb{R}$. We assume there exist $M_1 \in \mathbb{R}$, $m_1 \in \mathbb{R}_+$, $C_0 \in \mathbb{R}$ and a non-decreasing function $\eta : \mathbb{R}_+ \to \mathbb{R}_+$ such that for all $i \in \{1, \dots, M\}$, we have*

$$\begin{aligned} |f^a(0, v_i)| &\leq M_1, \\ |f^b(0, u)| &\leq M_1 + m_1|u|, \\ C_0 \leq \partial_y f^a(y, v_i) &\leq M_1\eta(|y|), \\ C_0 \leq \partial_y f^b(y, u) &\leq (M_1 + m_1|u|)\eta(|y|). \end{aligned}$$

As noted before, we are able to deduce existence and uniqueness of the IVP (3.4) from the assumptions introduced above.

Theorem 3.12 ([93, Thm 3.1])**.** *Let Assumptions 3.10 and 3.11 hold. Let $(u, \boldsymbol{\alpha}) =: z \in L^q(\Omega_T, \mathbb{R}^{1+M})$ and $y_0 \in L^\infty(\Omega)$. Then, the IVP (3.4) admits a unique weak solution $y \in \mathcal{W} \cap L^\infty(\Omega_T)$, which additionally satisfies $y \in C(\bar{\Omega}_{\varepsilon,T})$ for every $\varepsilon > 0$. Furthermore, the estimates*

$$\|y\|_{L^\infty(\Omega_T)} \leq C_1(\|z\|_{L^q(\Omega_T, \mathbb{R}^{1+M})} + \|y_0\|_{L^\infty(\Omega)} + 1)$$

and

$$\|y\|_{C(\bar{\Omega}_{\varepsilon,T})} \leq C_2(\varepsilon)(\|z\|_{L^q(\Omega_T, \mathbb{R}^{1+M})} + \|y_0\|_{L^2(\Omega)} + 1).$$

hold for some constants $C_1 > 0$ and $C_2(\varepsilon) > 0$.

Proof. The proof is deferred to Appendix B.6. □

[3] We acknowledge that the idea to analyze the case of distributed integer controls in time and space in a setting in which the operator in the corresponding abstract Cauchy problem exhibits maximal parabolic regularity is due to Christian Meyer, TU Dortmund.

As mentioned before, we give an additional assumption that enables us to prove the implication (c) in our main chain of approximation arguments (1.1).

Assumption 3.13 (Continuity properties of the nonlinearity)**.** *Let the superposition operator defined as $f_i^a(y)(t,x) := f_i^a(y(t,x))$ be a continuous mapping from $\mathcal{Y}$ to $L^2((0,T),\mathcal{H})$ for all $i \in \{1,\dots,M\}$ and let the superposition operator, defined as $f^b(y,u)(t,x) := f^b(y(t,x),u(t,x))$, be a continuous mapping from $\mathcal{Y}$ to the weak topology of $L^2((0,T),\mathcal{H})$.*

Again, we convexify the discrete-valued control and relax the mixed constraint by $\delta \geq 0$ to obtain the equivalent reformulation (case $\delta = 0$) and the feasibility relaxation (case $\delta > 0$) of (MIPEVO-TX). We make use of the structure asserted in Assumptions 3.10 and 3.11 and obtain

$$\begin{aligned}
\min_{y,u,\boldsymbol{\omega}} \; & J(y,u,\boldsymbol{v}) \\
\text{s.t. } & \partial_t y + Ay = -\textstyle\sum_{i=1}^M \boldsymbol{\omega}_i f_i^a(y,v_i) - f^b(y,u) \\
& y((0,\cdot)) = y_0 \\
& y|_{\partial\Omega} = 0 \\
& -\delta \leq \boldsymbol{\omega}_i(s) c(y(s),u(s),v_i) \text{ a.e. in } \Omega_T \text{ for } 1 \leq i \leq M \\
& \boldsymbol{v} = \textstyle\sum_{i=1}^M \boldsymbol{\omega}_i v_i \\
& \boldsymbol{\omega}(s) \in \mathbb{S}^M \text{ a.e. in } \Omega_T.
\end{aligned} \tag{$\mathsf{BC}_\delta^{\text{EVO-TX}}$}$$

Again, we state a proposition that establishes the desired equivalence and, as this is the most involved case, we present the argument in more detail than for the other cases.

Proposition 3.14. *Assume that Assumptions 3.10 and 3.11 hold. Let $V := \mathbb{R}^{n_v}$ for some $n_v \in \mathbb{N}$, $\{v_1,\dots,v_M\} \subset V$ and $\boldsymbol{v}$ be restricted to $L^\infty(\Omega_T,V)$. Then,* (MIPEVO-TX) *is equivalent to* ($\mathsf{BC}_0^{\text{EVO-TX}}$) *with the correspondence mapping*

$$\mathcal{F}_{(\mathsf{BC}_0^{\text{EVO-TX}})} \ni (y,u,\boldsymbol{\omega}) \mapsto \left(y,u,\sum_{i=1}^M \boldsymbol{\omega}_i v_i\right) \in \mathcal{F}_{(\text{MIPEVO-TX})}.$$

Proof. Firstly, we consider equivalence of the IVPs. As in the proof of Lemma 3.6 for the case of binary and relaxed controls that only vary over time, we note that the admissible $\boldsymbol{\omega}$ the IVP in ($\mathsf{BC}_\delta^{\text{EVO-TX}}$), in particular for $\delta = 0$, are a subset of the admissible $\boldsymbol{\alpha}$ in (3.4). Thus, existence and uniqueness of the solution of the state equation follows from Theorem 3.12.

Let $(y,u,\boldsymbol{\omega})$ be admissible for ($\mathsf{BC}_0^{\text{EVO-TX}}$). We know $u \in L^q(\Omega_T)$ and $\boldsymbol{\omega} \in L^\infty(\Omega_T,\mathbb{R}^M)$ by assumption. Then, Theorem 3.12 implies $y \in L^\infty(\Omega_T)$ and consequently, the estimates in Assumption 3.11 yield that the right hand side is an L^1-function. Thus, we can employ the identity (2.1) to obtain the equality

$$f(y,u,\boldsymbol{v}) = \sum_{i=1}^M f^a(y,v_i) + f^b(y,u)$$

and by uniqueness of the solution to the IVP also the correct solution of the IVP, which in turn yields a feasible point with the same objective value.

Let $(y, u, \boldsymbol{v})$ be admissible for (MIPEVO-TX). Then, there exists $\boldsymbol{\omega}$ satisfying $\boldsymbol{v} = \sum_{i=1}^{M} \boldsymbol{\omega}_i v_i$ by virtue of Carathéodory's theorem, see Theorem B.33. We note that $\boldsymbol{\omega}$ is not necessarily unique, but it always exists. However, for every such $\boldsymbol{\omega}$, the argument above holds true again, which gives the coincidence of the right hand sides of the IVPs and consequently, by virtue of Theorem 3.12, the coincidence of the solutions with y. Thus, all these points are feasible with coinciding objective values. □

Again, we derive the continuous relaxation by weakening the SOS1 property of $\boldsymbol{\omega}$ in ($\mathrm{BC}_0^{\text{EVO-TX}}$) to convex combinations and denote the new coefficients by $\boldsymbol{\alpha}$, which gives

$$
\begin{aligned}
\min_{y,u,\boldsymbol{\alpha}} \; & J(y,u,\boldsymbol{v}) \\
\text{s.t. } & \partial_t y + Ay = \textstyle\sum_{i=1}^{M} -\boldsymbol{\alpha}_i f_i^a(y, v_i) - f^b(y,u) \\
& y((0,\cdot)) = y_0 \\
& y|_{\partial\Omega} = 0 \\
& 0 \leq \boldsymbol{\alpha}_i(s) c(y(s), u(s), v_i) \text{ a.e. in } \Omega_T \text{ for } 1 \leq i \leq M \\
& \boldsymbol{v} = \textstyle\sum_{i=1}^{M} \boldsymbol{\alpha}_i v_i \\
& \boldsymbol{\alpha}(s) \in \operatorname{conv} \mathbb{S}^M \text{ a.e. in } \Omega_T.
\end{aligned}
\tag{$\mathrm{RC}^{\text{EVO-TX}}$}
$$

Thus, we have derived the equivalent convexified reformulations as well as the feasibility and continuous relaxations for three classes of (MIPDECO), for which we verify the approximation chain (1.1) in the remainder.

Remark 3.15. *For sake of simplicity, we work with homogeneous Dirichlet boundary conditions in Assumption 3.10 and the computational results in Chapter 10. However, different boundary conditions are possible as well. In this case, we need to take care that the embeddings in the Gelfand triple*

$$\mathcal{V} \hookrightarrow \mathcal{H} \cong \mathcal{H}^* \hookrightarrow \mathcal{V}^*$$

are compact. Usually, this requires that an appropriate extension operator exists, e.g. to leverage the embedding $H^1(\Omega) \hookrightarrow^c L^2(\Omega)$, see [94, Chap. 7]. *This can be handled by assuming the 1-extension property (see Definition B.34) for Ω in addition to Assumptions 3.7 and 3.10.*

Part II

The main approximation chain

Chapter 4

The abstract setting

In this chapter, we handle contribution (d). Although it can be regarded as a corollary of the other results, (d) is the contribution where we make statements about optimality, which is of course the overall aim. Furthermore, we motivate a desirable property for rounding algorithms in the presence of mixed constraints that depend on the discrete-valued controls. Therefore, we present it before the introduction of the tailored algorithm that is capable of satisfying it and the proofs for the steps (a), (b) and (c).

We use the reformulations and relaxations of (MIOCP) introduced in Section 2.1. Section 4.1 gives a brief approximation argument for this setting. In Section 4.2, we formalize the results concerning contribution (d) in Corollaries 4.5 to 4.7. Before starting, we make Assumption 4.1, which essentially postulates that the implication (c) holds in (1.1) by means of the concept of complete continuity.

Assumption 4.1. *Let Assumption 2.2 hold. Furthermore,*

1. *let the control-to-state operator $S_R : \mathcal{U} \times L^\infty(\Omega_T, \mathbb{R}^M) \to \mathcal{Y}$ be completely continuous (see Definition B.32) in the second argument,*

2. *let the mapping $J : \mathcal{Y} \times \mathcal{U} \times L^\infty(\Omega_T, V) \to \mathbb{R}$ be such that for all $u \in \mathcal{U}$, the mapping $(y, \boldsymbol{v}) \mapsto J(y, u, \boldsymbol{v})$ is jointly continuous from the product $(\mathcal{Y}, \|\cdot\|_\mathcal{Y}) \times (L^\infty(\Omega_T, V), \sigma(L^\infty, L^1))$ to $(\mathbb{R}, |\cdot|)$.*

4.1 Weak integer approximation for completely continuous solution mappings

The first lemma states that weak approximation of relaxed controls by a sequence of binary controls implies convergence of the corresponding sequence of objective values to the value of the relaxed objective under Assumption 4.1 .

Lemma 4.2. *Let Assumption 4.1 hold. Let $(y, u, \boldsymbol{\alpha}) \in \mathcal{F}_{(\mathrm{RC})}$. Assume there exists a sequence $(\boldsymbol{\omega}^{(n)})_n$ of binary controls that satisfies*

$$\sum_{i=1}^{M} \boldsymbol{\omega}_i^{(n)} v_i = \boldsymbol{v}^{(n)} \rightharpoonup^* \boldsymbol{v} = \sum_{i=1}^{M} \boldsymbol{\alpha}_i v_i \text{ in } L^\infty(\Omega_T, V).$$

Then, $J(y^{(n)}, u, \boldsymbol{v}^{(n)}) \to J(y, u, \boldsymbol{v})$ for $y^{(n)} \coloneqq S(u, \boldsymbol{v}^{(n)}) = S_R(u, \boldsymbol{\omega}^{(n)})$.

Proof. The claim follows immediately from the application of Assumption 4.1. □

If the vector $\boldsymbol{v}$ lives in an L^p-space on Ω_T, we can employ older analysis results to obtain the desired existence of such a sequence.

Theorem 4.3. *Let V satisfy Assumption 2.2. Let $\boldsymbol{\alpha}$ be a relaxed control. Then, there exists a sequence $(\boldsymbol{\omega}^{(n)})_n$ of binary controls such that the convergence*

$$\sum_{i=1}^{M} \boldsymbol{\omega}_i^{(n)} v_i = \boldsymbol{v}^{(n)} \rightharpoonup^* \boldsymbol{v} = \sum_{i=1}^{M} \boldsymbol{\alpha}_i v_i$$

holds in $L^p(\Omega_T, V)$ for $1 < p \leq \infty$ and the convergence

$$\sum_{i=1}^{M} \boldsymbol{\omega}_i^{(n)} v_i = \boldsymbol{v}^{(n)} \rightharpoonup \boldsymbol{v} = \sum_{i=1}^{M} \boldsymbol{\alpha}_i v_i$$

holds in $L^p(\Omega_T, V)$ for $1 \leq p < \infty$.

Proof. Assumption 2.2 enables the required L^1-L^∞-duality, see Theorem B.27. First, we note that it suffices to show the existence and weak* convergence result in $L^\infty(\Omega_T, V)$. A constructive proof that establishes a sequence with the desired properties can be found in Tartar's work [114, Thm 3 (2)]. The set K therein corresponds to our set $\{v_1, \dots, v_M\}$. To relate the sequence $(\boldsymbol{v}^{(n)})_n \subset L^\infty(\Omega_T, V)$ to the sequence $(\boldsymbol{\omega}^{(n)})_n$, we set

$$\boldsymbol{\omega}_i^{(n)} \coloneqq \chi_{(\boldsymbol{v}^{(n)})^{-1}(\{v_i\})},$$

where $(\boldsymbol{v}^{(n)})^{-1}$ denotes the preimage mapping of the function $\boldsymbol{v}^{(n)}$. □

We have cited a constructive proof. Existence results are even older. According to Tartar [114], they date *back essentially* to Lyapunov's article [78].[4] However, there exists a wealth of reinvestigations and extension of the result, known as *the Lyapunov convexity theorem*, and proof. A very short proof, which is based on the Krein-Milman theorem, is given by Lindenstrauss in [73]. In Yorke's article [121], a relationship to *bang-bang controls* for an ODE-based IVP is established. We also mention Artstein's concise summary that *the range of an atomless $\mathbb{R}^m$-valued measure is compact and convex* in the abstract of [7]. The works [7, 121] contain references to further applications and proofs of the result and similar findings. To relate these results to our work, we make the remark below.

[4] Unfortunately, the article [78] is inaccessible to the author as he is not familiar with the Russian language and has not found any translation.

Remark 4.4. *Emerging from Lindenstrauss' proof in* [73]*, the identity*

$$\left\{\int_{\Omega_T} fg\,\mathrm{d}\lambda : g \in L^\infty(\Omega_T), 0 \le g \le 1\right\} = \left\{\int_{\Omega_T} fg\,\mathrm{d}\lambda : g = \chi_K, K \in \mathcal{B}\right\}$$

holds for all test functions $f \in L^1(\Omega_T)$, i.e. we observe the, initially astonishing, fact that the extreme points of a compact, convex set can constitute a dense subset of the convex set itself in weaker topologies, in this case $(L^\infty(\Omega_T), \sigma(L^\infty, L^1))$, in infinite-dimensional vector spaces. In the remainder, results of this type link binary controls (extreme points) to relaxed controls (elements of the convex hull).

4.2 Optimality of the weak relaxed control approximation

In the absence of mixed constraints, we have $\mathcal{F}_{(\mathsf{BC}_0)} = \mathcal{F}_{(\mathsf{BC}_\delta)}$ and Lemma 4.2 and Theorem 4.3 imply the following consequences on the optimality of the binary-valued approximants. Variants of them are proven by the author in [79] for elliptic control systems.

Corollary 4.5 (Variant of [79, Thm 5.1])**.** *Let Assumption 4.1 hold. Let the assumptions of Theorem 4.3 hold, the constraint $0 \le c(y,u,\boldsymbol{v})$ be absent and* (RC) *admit a global minimizer. Then,*

$$\min_{(y,u,\boldsymbol{\alpha})\in\mathcal{F}_{(\mathsf{RC})}} J(y,u,\boldsymbol{\alpha}) = \inf_{(y,u,\boldsymbol{\omega})\in\mathcal{F}_{(\mathsf{BC}_0)}} J(y,u,\boldsymbol{\omega}).$$

Sometimes it may be realistic to assume a Tikhonov regularizer of the discrete-valued control, e.g. to simplify or accelerate the solution of (RC). As such terms are only norm continuous, but not weakly continuous, we obtain a suboptimality in this case.

Corollary 4.6 (Variant of [79, Cor. 5.2])**.** *Let the assumptions of Theorem 4.3 hold, the constraint $0 \le c(y,u,\boldsymbol{v})$ be absent and the objective of* (RC) *have the additional summand $\gamma\|\sum_{i=1}^M v_i\boldsymbol{\alpha}_i\|_{L^p}^p$. Assume that this modification of* (RC) *admits a global minimizer $(y^*,u^*,\boldsymbol{\alpha}^*)$. Then, there exists a sequence $(\boldsymbol{\omega}^{(n)}, y^{(n)})_n \subseteq L^\infty(\Omega_T)\times\mathcal{Y}$ with $(y^{(n)},u^*,\boldsymbol{\omega}^{(n)})_n \subset \mathcal{F}_{(\mathsf{BC}_0)}$ such that*

$$\begin{aligned}\lim J(y^{(n)},u^*,\boldsymbol{\omega}^{(n)}) &+ \gamma\|\boldsymbol{v}^{(n)}\|_{L^p}^p \\ &\le \inf_{(y,u,\boldsymbol{\omega})\in\mathcal{F}_{(\mathsf{BC}_0)}} J(y,u,\boldsymbol{\omega}) + \gamma\lambda(\Omega_T)\max_{i\in\{1,\dots,M\}}\|v_i\|_V^p\end{aligned}$$

for $\gamma > 0$ and $1 \le p < \infty$.

Proof. The proof leans on [79, Cor. 5.2]. Hölder's inequality and the properties of relaxed controls yield

$$\gamma\left\|\sum_{i=1}^M \boldsymbol{\alpha}_i^* v_i\right\|_{L^p}^p \le \gamma\lambda(\Omega_T)\max_{i\in\{1,\dots,M\}}\|v_i\|_V^p =: C.$$

We denote $\boldsymbol{v}^* \coloneqq \sum_{i=1}^M \boldsymbol{\alpha}_i^* v_i$ and the assumptions assert that the equality

$$J(y^*, u^*, \boldsymbol{\alpha}^*) + \gamma \|\boldsymbol{v}^*\|_{L^p}^p = \min_{(y,u,\boldsymbol{\alpha}) \in \mathcal{F}_{(\mathsf{RC})}} J(y, u, \boldsymbol{\alpha}) + \gamma \left\| \sum_{i=1}^M \boldsymbol{\alpha}_i v_i \right\|_{L^p}^p$$

holds. This gives the estimate

$$\inf_{(y,u,\boldsymbol{\alpha}) \in \mathcal{F}_{(\mathsf{RC})}} J(y, u, \boldsymbol{\alpha}) \geq J(y^*, u^*, \boldsymbol{\alpha}^*) + \gamma \left\| \sum_{i=1}^M \boldsymbol{\alpha}_i^* v_i \right\|_{L^p}^p - C.$$

By Theorem 4.3, there exists a sequence $(y^{(n)}, \boldsymbol{\omega}^{(n)})_n$ such that $J(y^{(n)}, u^*, \boldsymbol{\omega}^{(n)}) \to J(y^*, u^*, \boldsymbol{\alpha}^*)$, which implies

$$J(y^*, u^*, \boldsymbol{\alpha}^*) + \frac{\gamma}{2} \left\| \sum_{i=1}^M \boldsymbol{\alpha}_i^* v_i \right\|_{L^p}^p - C \geq \lim J(y^{(n)}, u^*, \boldsymbol{\omega}^{(n)}) - C.$$

We combine these inequalities with the fact that (RC) relaxes (BC_0) and deduce

$$\inf_{(y,u,\boldsymbol{\omega}) \in \mathcal{F}_{(\mathsf{BC}_0)}} J(y, u, \boldsymbol{\omega}) + C \geq \lim J(y^{(n)}, u^*, \boldsymbol{\omega}^{(n)}) \geq \inf_{(y,u,\boldsymbol{\omega}) \in \mathcal{F}_{(\mathsf{BC}_0)}} J(y, u, \boldsymbol{\omega}).$$

□

In the presence of the mixed constraint, we obtain a weaker statement.

Corollary 4.7. *Let the assumptions of Theorem 4.3 hold and* (RC) *admit a global minimizer* $(y^*, u^*, \boldsymbol{\alpha}^*)$*. Then, there exists a sequence* $(y^{(n)}, \boldsymbol{\omega}^{(n)})_n$ *such that*

$$J(y^{(n)}, u^*, \boldsymbol{\omega}^{(n)}) \to \min_{(y,u,\boldsymbol{\alpha}) \in \mathcal{F}_{(\mathsf{RC})}} J(y, u, \boldsymbol{\alpha}).$$

Let additionally $0 \leq c(y^*(s), u^*(s), \sum_{i=1}^M \boldsymbol{\alpha}_i^*(s) v_i)$ *hold for a.a.* $s \in \Omega_T$ *and the mapping* $(y, \boldsymbol{v}) \mapsto c(y, u^*, \boldsymbol{v})$ *be a jointly continuous mapping from the product* $(\mathcal{Y}, \|\cdot\|_{\mathcal{Y}}) \times (L^\infty(\Omega_T, V), \sigma(L^\infty, L^1))$ *to* $L^\infty(\Omega_T, \mathbb{R}^{n_c})$*. Then,*

$$0 \leq \liminf c\left(y^{(n)}, u^*, \sum_{i=1}^M \boldsymbol{\omega}_i^{(n)} v_i\right) \text{ in } L^\infty(\Omega_T, \mathbb{R}^{n_c}) \tag{4.1}$$

Proof. The first claim follows from Theorem 4.3 and the second from the assumed continuity properties of c. □

The statement is weaker because we cannot ensure that the pointwise a.e. constraint $0 \leq c(y^{(n)}(s), u^*(s), \sum_{i=1}^M \boldsymbol{\omega}_i^{(n)}(s) v_i)$ in (BC_0) holds without additional assumptions. Allowing the infeasibility in (4.1) is necessary and it is possible that a gap between the infimal values of (RC) and (BC_0) exists. We refer to an example by Cesari in [19, Chap. 18.7], which was investigated by Lenders in [68, 71].

Thus, we desire rounding algorithms that not only satisfy $\boldsymbol{\omega}^{(n)} \rightharpoonup^* \boldsymbol{\alpha}$ but also (4.1) when applied to solutions of (RC), i.e. when combined with the satisfaction of

$$0 \leq \boldsymbol{\alpha}_i(s) c(y(s), u(s), v_i) \text{ for a.a. } s \in \Omega_T \text{ for all } i \in \{1, \dots, M\}$$

for some relaxed control and corresponding state. In Sections 5.3 and 6.4, we introduce and analyze a variant of the SUR family that it is capable of doing so.

Chapter 5

A generalized Sum-Up Rounding algorithm

This chapter defines a class of SUR algorithms that unifies the generalizations from [68, 79, 82]. The algorithm is reformulated to operate on partitions of multi-dimensional domains instead of intervals. The set of admissible indices for rounding can be restricted on each discretization cell. This allows to derive a variant that satisfies the demands from Section 4.2 in the presence of pointwise mixed constraints involving the discrete controls in (MIOCP). Many of the following considerations closely follow the publications [68, 82] of Kirches, Lenders and the author and parts of the notation have been adopted from [79, 80] of Kirches and the author.

For refined rounding grids, the presented algorithms produce the sequences of binary controls that start the approximation chain (1.1). As we frequently use sequences of them, we denote them by $\boldsymbol{\omega}^{(n)}$ with n indexing the rounding grids. They correspond to the $\boldsymbol{\omega}^{(h)}$ in the approximation chain (1.1) if h denotes the maximum cell volume (mesh size) of the n-th rounding grid.

We begin with the notation to handle multi-dimensional domains and assign names to the recurring quantities in the definition and analysis of the algorithm. Then, we define the algorithm and put it in the context of our literature overview in Section 2.4. We point out why its standard variant might fail in the presence of the aforementioned constraints and state a variant that does not. Finally, we demonstrate that an analogous modification of the alternative algorithm Next-Forced Rounding (NFR), see also Section 2.4, fails.

5.1 Preparatory definitions

Unfortunately, there is no immediate or natural (continuous) analog to forward progression in time on $[0, T]$ when considering a multi-dimensional domain, or space-time cylinder, $\Omega_T \subset \mathbb{R}^d$. However, partitions of $\bar{\Omega}_T$ into N cells will serve our purpose.

Definition 5.1 (Grid partitioning $\bar{\Omega}_T$)**.** *Let $\Omega_T \subset \mathbb{R}^d$ be a bounded domain. We call the set $\{\mathcal{T}_1, \dots, \mathcal{T}_N\}$ a **rounding grid for** Ω_T if it is a finite partition of $\bar{\Omega}_T$, i.e. if $\mathcal{T}_k \in \mathcal{B}(\bar{\Omega}_T)$ for all k, $\bigcup_{k=1}^N \mathcal{T}_k = \bar{\Omega}_T$ and $\mathcal{T}_k \cap \mathcal{T}_\ell = \emptyset$ for $k \neq \ell$. We define*

$$\overline{\Delta} \coloneqq \max_{k \in \{1,\dots,N\}} \lambda(\mathcal{T}_k) \quad \text{and} \quad \underline{\Delta} \coloneqq \min_{k \in \{1,\dots,N\}} \lambda(\mathcal{T}_k)$$

*and call $\overline{\Delta}$ the **mesh size** of the grid.*

Having Definition 5.1 at hand, we take a rounding grid-based view and consider (sets of) grid cells in most of the remainder. For convenience of the informed reader who is familiar with the literature, we sometimes also state the continuous versions of the quantities for the one-dimensional case.

The so-called *control deviation* and *integrated control deviation* have already been mentioned in Section 2.4 and occur in our definition of the SUR algorithm. The integrated control deviation is of central interest to us. The bounds we establish for it in the next chapter constitute the first step in our approximation arguments.

Definition 5.2 (Control deviation)**.** *Let a bounded domain $\Omega_T \subset \mathbb{R}^d$ and let a relaxed control $\boldsymbol{\alpha}$ and a binary control $\boldsymbol{\omega}$ be given. The function*

$$\boldsymbol{\phi} \coloneqq \boldsymbol{\alpha} - \boldsymbol{\omega}$$

*is called **control deviation**. The function*

$$\boldsymbol{\Phi} : \mathcal{B}(\bar{\Omega}_T) \ni A \mapsto \int_A \boldsymbol{\phi}(s) \, \mathrm{d}s \in \mathbb{R}^M$$

*is called **integrated control deviation**. In case $\Omega_T = (0, T)$, we also call*

$$\boldsymbol{\Phi}_{\mathrm{1D}} : [0, T] \ni t \mapsto \int_0^t \boldsymbol{\alpha}(s) - \boldsymbol{\omega}(s) \, \mathrm{d}s \in \mathbb{R}^M$$

*the **integrated control deviation**.*

In a nutshell, the *control deviation* is the (sign-sensitive and vector-valued) difference between relaxed and binary control trajectories and the *integrated control deviation* is the integral of this quantity from zero to a point in time or over a subset of the grid cells in the multi-dimensional case.

5.2 The algorithm (SUR-GEN)

We generalize the SUR algorithm, cf. Definition 2.8. We give a definition from a function space point of view, i.e. the algorithm transforms an L^∞ function into another one. For the sake of brevity, the part *SOS* that has been used in the names of SUR algorithms in the literature is dropped as all considered algorithms secure the SOS1 property of the output function.

Definition 5.3 (Sum-Up Rounding Algorithms)**.** *Let $\{\mathcal{T}_1, \dots, \mathcal{T}_N\}$ be a rounding grid of a bounded domain Ω_T. For a relaxed control $\boldsymbol{\alpha}$ and sets $\emptyset \neq F_k \subset \{1, \dots, M\}$ for $k \in \{1, \dots, N\}$, we define the class of* ***generalized Sum-Up Rounding Algorithms*** *recursively by*

$$\begin{aligned} \boldsymbol{\omega}(s) &\coloneqq \sum_{k=1}^{N} \chi_{\mathcal{T}_k}(s) W_k, \\ W_k(i) &\coloneqq \begin{cases} 1 & \text{if } i = \underset{j \in F_k}{\arg\max} \int_{\mathcal{T}_k} \boldsymbol{\alpha}_j(s) \, \mathrm{d}s - \boldsymbol{\Phi}_j \left(\bigcup_{k'=1}^{k-1} \mathcal{T}_{k'} \right), \\ 0 & \text{else} \end{cases} \end{aligned} \tag{SUR-GEN}$$

for $i \in \{1, \dots, M\}$. The operation $\arg\max$ is implemented such that ambiguity of the maximum implies that exactly one of the maximizing indices is returned (e.g. the smallest) in a deterministic manner and $i \notin F_k$ implies $W_k(i) = 0$.

(SUR-GEN) is used to refer to this class of algorithms or any of them with unspecified sets F_k. We make the following proposition concerning well-definedness and computational complexity.

Proposition 5.4. *Let $\boldsymbol{\omega}$ be computed by* (SUR-GEN) *executed on a relaxed control $\boldsymbol{\alpha}$ under satisfaction of the assumptions of Definition 5.3. Then,*

1. *the behavior of* (SUR-GEN) *is well-defined,*
2. *$\boldsymbol{\omega}$ is a binary control,*
3. (SUR-GEN) *exhibits a computational complexity of $\mathcal{O}(N)$.*

Proof. The first claim holds true because the integrated control deviation $\boldsymbol{\Phi}$ only evaluates the integral on sets for which $\boldsymbol{\omega}$ has already been defined in previous iterations. The second and third claim follow by inspection. □

The choice of the admissible rounding indices F_k for a grid cell k is a degree of freedom for the algorithm. The following choice gives the standard variant, which works in the absence of pointwise a.e. mixed constraints that involve the discrete controls.

Definition 5.5 (SUR)**.** *We define the* Standard Sum-Up Rounding Algorithm (SUR) *as* (SUR-GEN) *with the setting*

$$F_k = \{1, \dots, M\} \tag{SUR}$$

for all grid cells $k \in \{1, \dots, N\}$.

We refer to this variant as (SUR). Most of the existing literature focuses on (SUR) in the one-dimensional case, which was also the originally developed variant of (SUR-GEN). The following example relates the discrete variant from Section 2.4 to (SUR-GEN).

Example 5.6. *We derive the quantities for* (SUR-GEN) *and Definition 5.1 for the one-dimensional time-dependent case. Let* $\bar{\Omega}_T \coloneqq [0,T]$ *consider the rounding grid induced by the nodes* $0 = t_0 < \ldots < t_N = T$*, i.e.* $\mathcal{T}_k \coloneqq [t_{k-1}, t_k)$ *for* $k \in \{1, \ldots, N-1\}$ *and* $\mathcal{T}_N \coloneqq [t_{N-1}, t_N]$*. Then,* (SUR) *coincides with the original algorithm* (SUR-SOS1)*, see* [97] *and Definition 2.8. The discrete functions therein can be obtained by setting*

$$\boldsymbol{\alpha}_i(k) \coloneqq \frac{1}{\lambda(\mathcal{T}_k)} \int_{\mathcal{T}_k} \boldsymbol{\alpha}_i(s)\,\mathrm{d}s \quad \textit{and} \quad \boldsymbol{\omega}_i(k) \coloneqq \frac{1}{\lambda(\mathcal{T}_k)} \int_{\mathcal{T}_k} \boldsymbol{\omega}_i(s)\,\mathrm{d}s.$$

5.3 A variant of (SUR-GEN) in the presence of pointwise mixed constraints

We recall that our approximation arguments will provide means to deduce the convergence $y^{(n)} \to y$ with y solving the state equation for $\boldsymbol{\alpha}$ and the $y^{(n)}$ solving the state equation for $\boldsymbol{\omega}^{(n)}$, which are the outputs of (SUR) for $\boldsymbol{\alpha}$ on a sequence of successively refined rounding grids. Consider a pointwise a.e. defined mixed constraint $0 \leq c(y, u, \boldsymbol{v})$, which is continuous with respect to y. Its outer convexification reads

$$0 \leq \boldsymbol{\omega}_i(s) c(y(s), u(s), v_i) \text{ for a.a. } s \in \Omega_T \text{ for all } i \in \{1, \ldots, M\},$$

which is an equivalent reformulation by virtue of Proposition 5.4 and the considerations in Section 2.1. The corresponding constraint in the continuous relaxation reads

$$0 \leq \boldsymbol{\alpha}_i(s) c(y(s), u, v_i) \text{ for a.a. } s \in \Omega_T \text{ for all } i \in \{1, \ldots, M\}.$$

Assume we are using (SUR). For the entry i with $\boldsymbol{\omega}_i^{(n)} = 1$ on some $\mathcal{T}_k^{(n)}$, we desire $0 \leq c(y^{(n)}, u, v_i)$ a.e. or at least only an arbitrarily small violation of the constraint on $\mathcal{T}_k^{(n)}$ for n sufficiently large. Unfortunately, the rounding rule of (SUR) does not prevent $\boldsymbol{\omega}_i^{(n)} = 1$ on $\mathcal{T}_k^{(n)}$ if $\boldsymbol{\alpha}_i(s) = 0$ for a.a. $s \in \mathcal{T}_k^{(n)}$. But in this case,

$$(\boldsymbol{\omega}_i^{(n)} c(y^{(n)}, u, v_i))|_{\mathcal{T}_k^{(n)}} = c(y, u, v_i)|_{\mathcal{T}_k^{(n)}}$$

may assume arbitrarily large negative values on a subset of $\mathcal{T}_k^{(n)}$ of positive measure. This severe constraint violation can be circumvented by modifying (SUR-GEN) as in [68]. To this end, consider the proposition below.

Proposition 5.7. *Let* $i \in \{1, \ldots, M\}$*,* $u \in \mathcal{U}$*, let the mapping* $\mathcal{Y} \ni y \mapsto c(y, u, v_i) \in L^\infty(\Omega_T)$ *be continuous,* $y^{(n)} \to y$ *in* $\mathcal{Y}$ *and* $0 \leq c(y, u, v_i)$ *hold a.e. in* Ω_T*. Then,*

$$0 \leq \liminf c(y^{(n)}(s), u(s), v_i) \text{ for a.a. } s \in \Omega_T.$$

Proof. The claim follows from the continuity of the mapping $y \mapsto c(y, u, v_i)$. □

Assume that we restrict the rounding in each cell k to the entries i with $\int_{\mathcal{T}_k^{(n)}} \boldsymbol{\alpha}_i > 0$ and assume that $0 \leq \boldsymbol{\alpha}_i c(y, u, v_i)$ is satisfied. If we obtain $0 \leq \liminf c(y^{(n)}, u, v_i)$ from Proposition 5.7 for these combinations of cells $\mathcal{T}_k^{(n)}$ and entries i, this complies with $\boldsymbol{\omega}_i = 1$ on $\mathcal{T}_k^{(n)}$. We formalize this restriction of the admissible indices for rounding in the following specialization of (SUR-GEN).

Definition 5.8 (SUR for Vanishing Constraints)**.** *The **Vanishing-Constraint Sum-Up Rounding Algorithm** is defined as* (SUR-GEN) *with*

$$F_k \coloneqq \left\{ i \in \{1, \dots, M\} : \int_{\mathcal{T}_k} \boldsymbol{\alpha}_i(s) \, \mathrm{d}s > 0 \right\} \tag{SUR-VC}$$

for all grid cells $k \in \{1, \dots, N\}$.

Inspecting Proposition 5.7, we observe the need for continuity in the first argument of the superposition mapping c to obtain feasible points for (BC_δ) for $\delta \to 0$, i.e. $0 \leq \liminf \boldsymbol{\omega}_i^{(n)} c(y^{(n)}, u, v_i)$. We formalize our considerations in Proposition 5.9, which states that feasible points of (RC) can be approximated arbitrarily well by feasible points of (BC_δ) with $\delta \to 0$ when using (SUR-VC) and assuming (1.1) holds. The corner stone for this, implication (a) for (SUR-VC), will be proven as Theorem 6.20 in Chapter 6.

Proposition 5.9. *Let Assumption 2.2 hold and* (1.1) *hold for* (SUR-VC) *on a sequence of partitions of Ω_T. Let $\boldsymbol{\omega}^{(n)}$ be computed by* (SUR-VC) *on the partition n from a relaxed control $\boldsymbol{\alpha}$. Let $u \in \mathcal{U}$. Let $y^{(n)} \coloneqq S_R(u, \boldsymbol{\omega}^{(n)})$ for $n \in \mathbb{N}$ and $y \coloneqq S_R(u, \boldsymbol{\alpha})$. Let $0 \leq \boldsymbol{\alpha}_i c(y, u, v_i)$ a.e. on Ω_T for some $i \in \{1, \dots, M\}$. Then,*

$$-\delta^{(n)} \leq \boldsymbol{\omega}_i^{(n)} c(y^{(n)}, u, v_i) \text{ a.e. in } \Omega_T$$

and $\delta^{(n)} \to 0$.

Proof. For $n \in \mathbb{N}$, we consider the cells $\mathcal{T}_k^{(n)}$ for $k \in \{1, \dots, N^{(n)}\}$ individually. Let $\boldsymbol{\omega}_i^{(n)}|_{\mathcal{T}_k^{(n)}} = 0$. Then, $\boldsymbol{\omega}_i^{(n)} c(y^{(n)}, u, v_i) = 0$ on $\mathcal{T}_k^{(n)}$. Thus, it suffices to consider the cases where $\boldsymbol{\omega}_i^{(n)}|_{\mathcal{T}_k^{(n)}} = 1$. By definition of (SUR-VC), this implies $\boldsymbol{\alpha}_i > 0$ on $\mathcal{T}_k^{(n)} \cap A$ with $\lambda(\mathcal{T}_k^{(n)} \cap A) > 0$ for the choice $A = \{s \in \Omega_T : \boldsymbol{\alpha}_i(s) > 0\}$. We can deduce $0 \leq c(y, u, v_i)$ a.e. in A from the prerequisites and obtain

$$(\boldsymbol{\omega}_i^{(n)} c(y^{(n)}, u, v_i))(s) = (c(y^{(n)}, u, v_i))(s) \geq -\delta_k^{(n)}$$

for $s \in \mathcal{T}_k^{(n)} \cap A$ and $\delta_k^{(n)} \to 0$ by Proposition 5.7. Choosing $\delta^{(n)} \coloneqq \max\{\delta_k^{(n)} : k \in \{1, \dots, N^{(n)}\}\}$ yields the claim. □

5.4 NFR in the presence of mixed constraints

In Section 2.4, we have mentioned NFR as a possible alternative to (SUR), which offers an improved bound on the control deviation by cost of quadratic instead of linear

effort for the rounding computation. Naturally, one may ask if a restriction of admissible rounding indices to the set F_k also works for NFR, which would mean that the algorithm can also be applied in the presence of mixed constraints. We note the feature of NFR that it computes a set of indices, which are considered as admissible choices for rounding, specifically

$$\mathcal{A}_k = \left\{ i \in \{1, \dots, M\} : \int_{\bigcup_{\ell=1}^{k} \mathcal{T}_\ell} \boldsymbol{\alpha}_i - \int_{\bigcup_{\ell=1}^{k-1} \mathcal{T}_\ell} \boldsymbol{\omega}_i \geq -\overline{\Delta} + \lambda(\mathcal{T}_k) \right\}$$

for the k-th cell, cf. [64, Def. 4.6, Alg. 4.1].

Modifying the algorithm implies that the set of admissible indices for cell k, $\mathcal{A}_k$, is further restricted to indices that are also contained in the set F_k. In this case, it can happen that the set $\mathcal{A}_k \cap F_k$ is empty and no rounding decision is possible, see Table 5.1 for an example of this situation. Here, the admissible indices for interval 2 would be $\mathcal{A}_2 = \{3, 4, 5\}$ but $F_2 = \{1, 2\}$. Thus, the intersection is empty and no feasible rounding decision can be made. Thus, this modification is not possible for NFR without loosing its well-definedness.

Table 5.1: Weighted mean of relaxed controls $\boldsymbol{\alpha}$ and the resulting control deviations $\boldsymbol{\phi}$ after application of NFR where $\overline{\Delta} = \lambda(\mathcal{T}_1) = ... = \lambda(\mathcal{T}_N)$.

i	$\int_{\mathcal{T}_1} \boldsymbol{\alpha}_i$	$\int_{\mathcal{T}_2} \boldsymbol{\alpha}_i$	$\int_{\mathcal{T}_3} \boldsymbol{\alpha}_i$	$\int_{\mathcal{T}_1} \boldsymbol{\phi}_i$	$\int_{\bigcup_{k=1}^{2} \mathcal{T}_k} \boldsymbol{\phi}_i$	$\int_{\bigcup_{k=1}^{3} \mathcal{T}_k} \boldsymbol{\phi}_i$
1	$\frac{1}{5}$	$\frac{1}{4}$	$\frac{1}{2}$	$-\frac{4}{5}$	$-\frac{11}{20}$	$\lightning$
2	$\frac{1}{5}$	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{5}$	$-\frac{11}{20}$	$\lightning$
3	$\frac{1}{5}$	$\frac{1}{4}$	0	$\frac{1}{5}$	$\frac{9}{20}$	$\lightning$
4	$\frac{1}{5}$	$\frac{1}{4}$	0	$\frac{1}{5}$	$\frac{9}{20}$	$\lightning$
5	$\frac{1}{5}$	0	0	$\frac{1}{5}$	$\frac{4}{20}$	$\lightning$

Chapter 6

Approximation properties of (SUR-GEN)

This chapter serves to achieve contribution (a) of (1.1), i.e. we show

$$\overline{\Delta}^{(n)} \to 0 \quad \underset{(a)}{\Rightarrow} \quad d^{(n)}(\boldsymbol{\omega}^{(n)}, \boldsymbol{\alpha}) \to 0$$

for the two variants (SUR) and (SUR-VC) of (SUR-GEN). Here, $\boldsymbol{\alpha}$ is a relaxed control, $(\boldsymbol{\omega}^{(n)})_n$ is the sequence of binary controls produced by (SUR-GEN) on a sequence of rounding grids and $(\overline{\Delta}^{(n)})_n$ is the corresponding sequence of mesh sizes. We begin with the definition of the $d^{(n)}$ and introduce a name for the convergence $d^{(n)}(\boldsymbol{\omega}^{(n)}, \boldsymbol{\alpha}) \to 0$ as well as the notation for the subsequent sections. Then, we prove the bound

$$d(\boldsymbol{\omega}, \boldsymbol{\alpha}) \leq \mathcal{O}(\log(M))\overline{\Delta}$$

for (SUR). This bound was conjectured in [100, Rem. 7] and is proven in [68]. We offer a novel proof in Section 6.3 that also works for non-equidistant rounding grids. Subsequently, we analyze the algorithm (SUR-VC) and provide reworks of the author's proofs from [68, 82]. First, we prove the bound

$$d(\boldsymbol{\omega}, \boldsymbol{\alpha}) \leq (M+1)\overline{\Delta}$$

and then, we improve it to

$$d(\boldsymbol{\omega}, \boldsymbol{\alpha}) \leq \left\lfloor \frac{M}{2} \right\rfloor \overline{\Delta},$$

which is far more technical to prove but asymptotically tight because an example given in [68] reaches this bound. Finally, we demonstrate that we cannot expect that all variants of (SUR-GEN) behave nicely w.r.t. the mesh size of the rounding grid.

6.1 The integrality gap

We define the notion of *integrality gap* formally.

Definition 6.1 (Integrality gap)**.** *Let the notation of Definition 5.2 hold. Let $\{\mathcal{T}_1, \dots, \mathcal{T}_N\}$ be a rounding grid of a bounded domain $\Omega_T \subset \mathbb{R}^d$. Then, we define the **integrality gap** between $\boldsymbol{\alpha}$ and $\boldsymbol{\omega}$ as*

$$d(\boldsymbol{\omega}, \boldsymbol{\alpha}) \coloneqq \max_{k \in \{1,\dots,N\}} \left\| \boldsymbol{\Phi}\left(\bigcup_{k'=1}^{k} \mathcal{T}_{k'} \right) \right\|_{\infty}.$$

If $\Omega_T = (0,T)$, we also denote by

$$d_{\mathrm{1D}}(\boldsymbol{\omega}, \boldsymbol{\alpha}) \coloneqq \operatorname*{ess\,sup}_{t \in [0,T]} \|\boldsymbol{\Phi}(t)\|_{\infty}$$

*the **integrality gap** between $\boldsymbol{\alpha}$ and $\boldsymbol{\omega}$.*

The $\operatorname{ess\,sup}$ operation for $\Omega_T = (0,T)$ emphasizes that our definition is designed to measure $\boldsymbol{\Phi}$ in the $\|\cdot\|_{L^\infty}$-norm. However, the $\max$ operation would also work because the pointwise interpretation of $\boldsymbol{\Phi}_{\mathrm{1D}}$ in Definition 5.2 gives the continuous representative of $\boldsymbol{\Phi}_{\mathrm{1D}} \in W^{1,\infty}((0,T), \mathbb{R}^M) \hookrightarrow C([0,T], \mathbb{R}^M)$.

Proposition 6.2. *Let $\boldsymbol{\omega}$ be computed by* (SUR-GEN) *from a relaxed control $\boldsymbol{\alpha}$ on a rounding grid that consists of subsequent intervals. Then, $d_{\mathrm{1D}}(\boldsymbol{\omega}, \boldsymbol{\alpha}) = d(\boldsymbol{\omega}, \boldsymbol{\alpha})$.*

Proof. The discretization is $0 < t_1 < \ldots < T_N = T$. The function$\boldsymbol{\Phi}_{\mathrm{1D}}$ assumes its maximum over the intervals $\bar{\mathcal{T}}_k = [t_k, t_{k+1}]$ at the interval boundaries, see [68, Lem. 5.7]. Thus, the definition of $d_{\mathrm{1D}}(\boldsymbol{\omega}, \boldsymbol{\alpha})$ is consistent with $d(\boldsymbol{\omega}, \boldsymbol{\alpha})$ in this case. □

We define the *vanishing integrality gap* formally.

Definition 6.3 (Vanishing integrality gap)**.** *Let a sequence of rounding grids be given with its corresponding sequence of integrality gaps $(d^{(n)})_n$. Let $\boldsymbol{\alpha}$ be a relaxed control and $(\boldsymbol{\omega}^{(n)})_n$ be a sequence of binary controls. Let $d^{(n)}(\boldsymbol{\omega}^{(n)}, \boldsymbol{\alpha}) \to 0$. Then, we say that the sequence of corresponding control deviations $(\boldsymbol{\phi}^{(n)})_n$, see Definition 5.2, is of **vanishing integrality gap**.*

*Let $\Omega_T = (0,T)$. Let $\boldsymbol{\alpha}$ be a relaxed control and $(\boldsymbol{\omega}^{(n)})_n$ be a sequence of binary controls. Let $d_{\mathrm{1D}}(\boldsymbol{\omega}^{(n)}, \boldsymbol{\alpha}) \to 0$. Then, we also say that the sequence of corresponding control deviations $(\boldsymbol{\phi}^{(n)})_n$ is of **vanishing integrality gap**.*

In the remainder, $\boldsymbol{\alpha}$ usually denotes a relaxed control solving (RC) and $\boldsymbol{\omega}^{(n)} = \boldsymbol{\omega}^{(n)}(\boldsymbol{\alpha})$ the binary control computed by (SUR-GEN) from $\boldsymbol{\alpha}$ on the n-th rounding grid.

6.2 Preparations

In this section, we introduce notation and a problem reduction to prepare the proofs in the remainder of this chapter. The proofs take a rounding grid-based or discrete view on $\boldsymbol{\alpha}$, $\boldsymbol{\omega}(\boldsymbol{\alpha})$ and $\boldsymbol{\phi}(\boldsymbol{\alpha})$. We mostly abbreviate $\boldsymbol{\omega} \coloneqq \boldsymbol{\omega}(\boldsymbol{\alpha})$ and $\boldsymbol{\Phi} \coloneqq \boldsymbol{\Phi}(\boldsymbol{\alpha})$ in absence of possible ambiguities and tacitly assume that $\boldsymbol{\omega}$ is produced by the currently considered variant of (SUR-GEN) for a given $\boldsymbol{\alpha}$ and that $\boldsymbol{\Phi}$ is computed accordingly.

6.2.1 Reduction to a discrete vantage point

We define the discrete quantities below that relate our quantities of interest to the grid:

$$\begin{aligned} h_k &\coloneqq \lambda\left(\mathcal{T}_k\right), \\ \boldsymbol{\alpha}_{k,i} &\coloneqq \frac{1}{h_k}\int_{\mathcal{T}_k} \boldsymbol{\alpha}_i, \\ \boldsymbol{\omega}_{k,i} &\coloneqq \frac{1}{h_k}\int_{\mathcal{T}_k} \boldsymbol{\omega}_i, \\ \boldsymbol{\Phi}_{k,i} &\coloneqq \boldsymbol{\Phi}_i\left(\bigcup_{k'=1}^{k} \mathcal{T}_{k'}\right) = \int_{\bigcup_{k'=1}^{k}\mathcal{T}_{k'}} \boldsymbol{\alpha}_i - \boldsymbol{\omega}_i, \quad \boldsymbol{\Phi}_{0,i} \coloneqq 0, \\ \boldsymbol{\gamma}_{k,i} &\coloneqq \boldsymbol{\alpha}_{k,i} h_k + \boldsymbol{\Phi}_{k-1,i}. \end{aligned}$$

Thus, h_k is the volume of the k-th cell, $\boldsymbol{\alpha}_{k,i}$ is the i-th component of the mean of $\boldsymbol{\alpha}$ on the k-th cell, implying $\boldsymbol{\alpha}_{k,\cdot} \in \operatorname{conv}\mathbb{S}^M$, $\boldsymbol{\omega}_{k,i}$ is the i-th component of the mean of $\boldsymbol{\omega}$ on the k-th cell, implying $\boldsymbol{\omega}_{k,\cdot} \in \mathbb{S}^M$ and $\boldsymbol{\omega}_{k,i} = W_k(i)$ in (SUR-GEN), and $\boldsymbol{\Phi}_{k,i}$ is the i-th component of the integrated control deviation of the cells 1 to k. The symbol $\boldsymbol{\gamma}_{k,i}$ is the *one cell lookahead quantity* on which (SUR-GEN)'s rounding decision is predicated, i.e. $\boldsymbol{\gamma}_{k,i} = \boldsymbol{\Phi}_{k,i} + \boldsymbol{\omega}_{k,i} h_k$.

Let the $\max$ in Definition 6.1 be extended to countable increasing unions of cells. If a bound can be established for this extension, it holds for the increasing union of cells until a finite cell index as well. Thus, we continue the sequence $(\mathcal{T}_k)_{k\in\{1,\dots,N\}}$ to $(\mathcal{T}_k)_{k\in\mathbb{N}}$ by adding virtual cells of volume $\overline{\Delta}$ for $k > N$. In the one-dimensional case with a discretization into intervals, this means that we take the supremum over the time horizon extended to $[0,\infty)$. To state the approximation properties easily, we introduce the set of admissible relaxed controls in our rounding grid-based view as

$$\mathcal{A} \coloneqq \left\{(\boldsymbol{\alpha}_{k,\cdot})_{k\in\mathbb{N}} : \boldsymbol{\alpha}_{k,\cdot} \in \operatorname{conv}\mathbb{S}^M \text{ for all } k \in \mathbb{N}\right\}.$$

We summarize these considerations into the following proposition.

Proposition 6.4. *Let a variant of* (SUR-GEN) *be fixed. Then, its computation is well-defined for all $\boldsymbol{\alpha} \in \mathcal{A}$ and all $(h_k)_{k\in\mathbb{N}} \subset (0,\overline{\Delta}]$. If there exists $C > 0$ such that*

$$\sup_{k\in\mathbb{N}} \|\boldsymbol{\Phi}_{k,\cdot}(\boldsymbol{\alpha})\|_\infty \le C$$

for all $\boldsymbol{\alpha} \in \mathcal{A}$ and all $(h_k)_{k\in\mathbb{N}} \subset (0,\overline{\Delta}]$. Then, for all relaxed controls $\boldsymbol{\alpha}$ and rounding grids $\{\mathcal{T}_1,\dots,\mathcal{T}_N\}$ of a bounded domain Ω_T, we have

$$d(\boldsymbol{\omega}(\boldsymbol{\alpha}),\boldsymbol{\alpha}) \le C.$$

This still holds if additional requirements on $(h_k)_{k\in\mathbb{N}}$ are imposed, which need not affect finitely many initial elements.

Proof. With the definitions of the discrete quantities above, we see that for each relaxed control $\boldsymbol{\alpha}$ and each discretization $(\mathcal{T}_k)_{k\in\{1,\dots,N\}}$, we obtain $\boldsymbol{\alpha}_{1,\cdot},\dots,\boldsymbol{\alpha}_{N,\cdot} \in \operatorname{conv}\mathbb{S}^M$ and $h_1,\dots,h_N$. Furthermore, for $\boldsymbol{\alpha}_{1,\cdot},\dots,\boldsymbol{\alpha}_{N,\cdot} \in \operatorname{conv}\mathbb{S}^M$, we can define the (canonical) relaxed control $\tilde{\boldsymbol{\alpha}} \coloneqq \sum_{i=1}^N \chi_{\mathcal{T}_i}\boldsymbol{\alpha}_{i,\cdot}$.

The $\arg\max$ is the only nonlinear operation in (SUR-GEN) and the relaxed and binary controls enter the integrals always additively. Thus, the rounding decision for $\boldsymbol{\alpha}$ and $\tilde{\boldsymbol{\alpha}}$ coincide, i.e. the rounding decision is based on the discrete quantity $\boldsymbol{\alpha}_{\cdot,\cdot}$ and the cell volumes h_k only, which gives the desired well-definedness.

Assume some bound holds if $(\boldsymbol{\alpha}_{k,\cdot})_k$ and $(h_k)_k$ are extended to $k > N$. Then, the bound also holds for $k \leq N$. Finally, if additional properties on $(h_k)_{k\in\mathbb{N}}$ are requested such that the initial N elements are not affected, then the bound holds true for finitely many cells without any restrictions. □

With the canonical construction in the proof of Proposition 6.4, we can apply (SUR-GEN) to $\boldsymbol{\alpha} \in \mathcal{A}$ by just applying the algorithm over the iterations $k \in \mathbb{N}$ instead of $k = 1,\dots,N$. We supply several properties of $(\boldsymbol{\Phi}_{k,\cdot})_{k\in\mathbb{N}}$ for given $\boldsymbol{\alpha} \in \mathcal{A}$ and $(h_k)_{k\in\mathbb{N}}$ that can be derived from (SUR-GEN) straightforwardly. We frequently exploit the fact that, for the discrete vantage point, the domain and the specific rounding grid are irrelevant and the algorithm only requires $\boldsymbol{\alpha} \in \mathcal{A}$ and $(h_k)_{k\in\mathbb{N}}$ to be well-defined.

Lemma 6.5 ([82, Lem. 4.2]). *Let $\boldsymbol{\alpha} \in \mathcal{A}$ and $(h_k)_k \subset (0,\overline{\Delta}]$ for some $\overline{\Delta} > 0$. Let $\boldsymbol{\omega}_{\cdot,\cdot}$, $\boldsymbol{\Phi}_{\cdot,\cdot}$ be computed from $\boldsymbol{\alpha}$ by means of* (SUR-GEN). *Then, the following properties hold for all $k \in \mathbb{N}$:*

1. $\boldsymbol{\alpha}_{k,j} \geq 0$ *for all* $j \in \{1,\dots,M\}$,
2. $\sum_{j=1}^M \boldsymbol{\alpha}_{k,j} = 1$,
3. $\sum_{j=1}^M [\boldsymbol{\Phi}]^+_{k,j} = \sum_{j=1}^M [\boldsymbol{\Phi}]^-_{k,j}$,
4. $\boldsymbol{\Phi}_{k,i} \geq 0$ *for at least one* $i \in \{1,\dots,M\}$,
5. $\boldsymbol{\Phi}_{k,i} \leq 0$ *for at least one* $i \in \{1,\dots,M\}$.

Proof. The definitions of $\boldsymbol{\alpha}_{k,\cdot}$, $\boldsymbol{\Phi}_{k,\cdot}$ imply that the (relaxed) SOS1 property carries over to the discrete vantage point, from which the first two assertions follow. The definition of (SUR-GEN) gives that the positive and negative part of $\boldsymbol{\Phi}_{\cdot,\cdot}$ have to sum up to the same value on each cell, which implies 3.-5., see [68, 82, 97, 100]. □

6.2.2 Ordering of entries in the integrated control deviation vector

Our proofs require a distinction between (strictly) positive and (strictly) negative entries. Thus, we define I_k^+ and I_k^{++} as the sets of indices i corresponding to (strictly) positive

entries of $\mathbf{\Phi}_{k,\cdot}$ and I_k^- and I_k^{--} analogously for (strictly) negative entries of $\mathbf{\Phi}_{k,\cdot}$, i.e.

$$\begin{aligned} I_k^+ &\coloneqq \{i \in \{1,\dots,M\} : \mathbf{\Phi}_{k,i} \geq 0\}, \\ I_k^{++} &\coloneqq \{i \in \{1,\dots,M\} : \mathbf{\Phi}_{k,i} > 0\}, \\ I_k^- &\coloneqq \{i \in \{1,\dots,M\} : \mathbf{\Phi}_{k,i} \leq 0\}, \\ I_k^{--} &\coloneqq \{i \in \{1,\dots,M\} : \mathbf{\Phi}_{k,i} < 0\}. \end{aligned}$$

The succeeding arguments require an encoding of the order of the entries in the $\mathbf{\Phi}_{k,\cdot}$, which is provided in the following definition.

Definition 6.6 (Encoding of the order within $[\mathbf{\Phi}]_{k,\cdot}^+$, $[\mathbf{\Phi}]_{k,\cdot}^-$, [82, Def. 2.1])**.** *For a sign $\mathfrak{s} \in \{+,-\}$, we name the elements of $I_k^{\mathfrak{s}}$ as follows,*

$$I_k^{\mathfrak{s}} = \left\{ i_1^{k,\mathfrak{s}}, \dots, i_{|I_k^{\mathfrak{s}}|}^{k,\mathfrak{s}} \right\},$$

where the $i_j^{k,\mathfrak{s}}$ encode a non-ascending ordering of the magnitude of the entries of $[\mathbf{\Phi}]_{k,\cdot}^{\mathfrak{s}}$,

$$[\mathbf{\Phi}]_{k,i_j^{k,\mathfrak{s}}}^{\mathfrak{s}} \geq [\mathbf{\Phi}]_{k,i_{j+1}^{k,\mathfrak{s}}}^{\mathfrak{s}}$$

for all $j \in \{1,\dots,|I_k^{\mathfrak{s}}|-1\}$. For $j \in \{1,\dots,|I_k^{\mathfrak{s}}|\}$, we define

$$\boldsymbol{\psi}_{k,j}^{\mathfrak{s}} \coloneqq \sum_{\ell=1}^{j} e_{i_\ell^{k,\mathfrak{s}}} \mathbf{\Phi}_{k,i_\ell^{k,\mathfrak{s}}}$$

Let k_0, $k_1 \in \mathbb{N}$, $k_0 \leq k_1$, $\mathfrak{s} \in \{+,-\}$ and a subset $J \subset I_{k_0}^{\mathfrak{s}}$ be given. Furthermore, let $J \subset I_k^{\mathfrak{s}}$ for all $k \in \{k_0,\dots,k_1\}$. For $k \in \{k_0,\dots,k_1\}$, we assign names $J = \{j_1^k,\dots,j_{|J|}^k\}$ to the elements of J and again define the indices j_ℓ^k to encode a non-ascending ordering,

$$[\mathbf{\Phi}]_{k,j_\ell^k}^{\mathfrak{s}} \geq [\mathbf{\Phi}]_{k,j_{\ell+1}^k}^{\mathfrak{s}}$$

for $\ell < |J|$. Furthermore, we define

$$\boldsymbol{\psi}_{k,\ell}^{J} \coloneqq \sum_{m=1}^{\ell} e_{j_m^k} \mathbf{\Phi}_{k,j_m^k}$$

For later convergence limits $\left[\bar{\mathbf{\Phi}}\right]^{\mathfrak{s}}$, we define $\bar{j}_\ell$ by

$$\left[\bar{\mathbf{\Phi}}\right]_{\bar{j}_\ell}^{\mathfrak{s}} \geq \left[\bar{\mathbf{\Phi}}\right]_{\bar{j}_{\ell+1}}^{\mathfrak{s}}.$$

for $\ell < |J|$, $J \subset I_{k_0}^{\mathfrak{s}}$.

Thus, $i_j^{k,\mathfrak{s}}$ is *the index of the j-th largest entry in $[\mathbf{\Phi}]_{k,\cdot}^{\mathfrak{s}}$.* The vector $\boldsymbol{\psi}_{k,j}^{\mathfrak{s}} \in \mathbb{R}^m$ can be described as $\mathbf{\Phi}_{k,\cdot}$, *where all of the entries except for the j positive (or negative) ones with the largest values are set to 0.* Similarly to above, $j_\ell^{k,\mathfrak{s}}$ is *the index of the entry in $\mathbf{\Phi}_{k,\cdot}$, which is the ℓ-th largest among the values $\left\{[\mathbf{\Phi}]_{k,j}^{\mathfrak{s}} : j \in J\right\}$* and the vector $\boldsymbol{\psi}_{k,\ell}^{J} \in \mathbb{R}^m$ can be described as $\mathbf{\Phi}_{k,\cdot}$ *with all entries set to 0 except for the ℓ positive (or negative) ones in J with the largest values.*

6.2.3 (SUR-GEN) in construction algorithms

Several proofs require the existence of certain control deviations $\boldsymbol{\Phi}_{k,\cdot}$. This is proven by providing algorithms that construct $\boldsymbol{\alpha}_{k_1+1,\cdot},\dots,\boldsymbol{\alpha}_{k,\cdot} \in \operatorname{conv}\mathbb{S}^M$ starting from a given $\boldsymbol{\Phi}_{k_1,\cdot}$. Appending them to given $\boldsymbol{\alpha}_{1,\cdot},\dots,\boldsymbol{\alpha}_{k_1,\cdot} \in \operatorname{conv}\mathbb{S}^M$, the application of (SUR-GEN) yields the desired $\boldsymbol{\Phi}_{k,\cdot}$. We use lines of the form

$$\boldsymbol{\Phi}_{k+1,\cdot} \leftarrow \text{(SUR-GEN)}(\boldsymbol{\Phi}_{k,\cdot}, \boldsymbol{\alpha}_{k+1,\cdot}, h_{k+1}, \text{ROUNDING ON PARITY}),$$

to compute to the intermediate values $\boldsymbol{\Phi}_{k_1+1,\cdot},\dots,\boldsymbol{\Phi}_{k,\cdot}$ in the algorithms. We interpret them as the application of the rounding rule of (SUR-GEN) in the $k+1$-th cell for given integrated control deviation until the k-th cell $\boldsymbol{\Phi}_{k,\cdot}$, relaxed control in the $k+1$-th grid cell $\boldsymbol{\alpha}_{k+1,\cdot}$, and volume of the $k+1$-th cell h_{k+1}. The fourth argument determines the rule for resolving the case that $\arg\max$ in (SUR-GEN) is not single-valued. We omit the fourth argument if this cannot happen in the algorithm or does not affect the result.

6.3 A bound on the integrality gap for (SUR)

In this section, we establish the following theorem.

Theorem 6.7. *Let $\boldsymbol{\omega}$ be a binary control computed by* (SUR) *for a relaxed control $\boldsymbol{\alpha}$ on a rounding grid with mesh size $\overline{\Delta}$. Then,*

$$d(\boldsymbol{\omega}, \boldsymbol{\alpha}) \leq C \cdot \overline{\Delta}$$

holds with $C = \mathcal{O}(\log(M))$.

Outline of proof. We employ Proposition 6.4 and prove the estimate by verifying

$$\sup_{k\in\mathbb{N}} \|\boldsymbol{\Phi}_{k,\cdot}\|_\infty \leq C \cdot \overline{\Delta}$$

for all $\boldsymbol{\alpha} \in \mathcal{A}$ and $(h_k)_{k\in\mathbb{N}} \subset (0, \overline{\Delta}]$ with the aforementioned assumption that $h_k = \overline{\Delta}$ for $k \geq N$ if N denotes the number of cells that make up the rounding grid. For (SUR), the bound on $\| [\boldsymbol{\Phi}]^+_{k,\cdot} \|_\infty$ is much harder to establish than the bound on $\| [\boldsymbol{\Phi}]^-_{k,\cdot} \|_\infty$, which is our first step towards the proof. We establish the proof in three steps:

1. Derivation of a bound on $\| [\boldsymbol{\Phi}]^-_{k,\cdot} \|_1$ in Section 6.3.1.
2. Show that the bound on $\| [\boldsymbol{\Phi}]^-_{k,\cdot} \|_1$ is tight in Section 6.3.2.
3. Employing the bound to maximize $\| [\boldsymbol{\Phi}]^+_{k,\cdot} \|_\infty$ in Section 6.3.3.

6.3.1 A bound on the sum-norm of the integrated control deviation

For this subsection, let $\boldsymbol{\alpha} \in \mathcal{A}$ be fixed. The bound is stated formally in the following proposition.

Proposition 6.8. *Let $k \in \mathbb{N}$ and $m \in \{1, \dots, M\}$ such that $|I_k^-| \geq m$. Then,*

$$\left\| \boldsymbol{\psi}_{k,m}^- \right\|_1 \leq \sum_{i=0}^{m-1} \left(1 - \sum_{j=0}^{i} \frac{1}{M-j} \right) \overline{\Delta}. \tag{6.1}$$

Proof. The proof employs mathematical induction over m. The base case is elaborated in Lemma 6.9 and the step in Lemma 6.10, which are presented below. The latter makes use of a summation formula, which is given in the Appendix as Lemma A.1. □

The arguments for the base case and the induction step of Proposition 6.8 are similar. The base can be followed more easily and is therefore presented first.

Lemma 6.9 (Base case)**.** *Let $|I_k^-| \geq 1$. Then,*

$$[\boldsymbol{\Phi}]_{k,i_1^{k,-}}^- \leq \left(1 - \frac{1}{M} \right) \overline{\Delta}.$$

Proof. We proceed by contradiction and assume that a cell $k \in \mathbb{N}$ exists such that

$$[\boldsymbol{\Phi}]_{k,i_1^{k,-}}^- > \left(1 - \frac{1}{M} \right) \overline{\Delta}.$$

Without loss of generality, we choose k such that this happens for the first time. Then, we know that (SUR) caused $\boldsymbol{\Phi}_{k,i_1^{k,-}} < \boldsymbol{\Phi}_{k-1,i_1^{k,-}}$ implying

$$\boldsymbol{\gamma}_{k,i_1^{k,-}} = \boldsymbol{\Phi}_{k,i_1^{k,-}} + h_k \leq \boldsymbol{\Phi}_{k,i_1^{k,-}} + \overline{\Delta} < \frac{1}{M} \overline{\Delta}.$$

Furthermore, Lemma 6.5 implies

$$\sum_{i \in I_k^{++}} [\boldsymbol{\Phi}]_{k,i}^+ > \left(1 - \frac{1}{M} \right) \overline{\Delta},$$

which in turn implies that there exists an index $i \neq i_1^{k,-}$ such that

$$[\boldsymbol{\Phi}]_{k,i}^+ > \frac{1 - \frac{1}{M}}{|I_k^{++}|} \overline{\Delta} \geq \frac{1 - \frac{1}{M}}{M-1} \overline{\Delta} = \frac{1}{M} \overline{\Delta}.$$

Combining the derivations above with the fact that the $i_1^{k,-}$-th entry was reduced, we obtain

$$\boldsymbol{\gamma}_{k,i} = [\boldsymbol{\Phi}]_{k,i}^+ > \frac{1}{M} \overline{\Delta} > \boldsymbol{\gamma}_{k,i_1^{k,-}}$$

which contradicts the $\arg\max$ in (SUR). This closes the proof. □

In a nutshell, Lemma 6.9 states that the largest value of the negative entries of $\mathbf{\Phi}_{k,\cdot}$ is bounded by $\left(1-\frac{1}{M}\right)\overline{\Delta}$, which already proves a tighter bound for $\|\,[\mathbf{\Phi}]^-_{k,\cdot}\,\|_\infty$ than we get for $\|\mathbf{\Phi}_{k,\cdot}\|_\infty$. We essentially repeat the arguments, which just look a little more complicated, to prove the induction step from the induction hypothesis.

Lemma 6.10 (Step)**.** *Assume* $|I_k^-| \geq m-1$ *implies*

$$\left\|\boldsymbol{\psi}^-_{k,m-1}\right\|_1 \leq \sum_{i=0}^{m-2}\left(1-\sum_{j=0}^{i}\frac{1}{M-j}\right)\overline{\Delta}.$$

Then, $|I_k^-| \geq m$ *implies*

$$\left\|\boldsymbol{\psi}^-_{k,m}\right\|_1 \leq \sum_{i=0}^{m-1}\left(1-\sum_{j=0}^{i}\frac{1}{M-j}\right)\overline{\Delta}.$$

Proof. Let $|I_k^-| \geq m$. As in proof of the base case in Lemma 6.9, we assume that the claim is false yielding that there exists $k \in \mathbb{N}$ such that

$$\left\|\boldsymbol{\psi}^-_{k,m}\right\|_1 = \sum_{j=1}^{m}[\mathbf{\Phi}]^-_{k,i_j^{k,-}} > \sum_{i=0}^{m-1}\left(1-\sum_{j=0}^{i}\frac{1}{M-j}\right)\overline{\Delta}$$

for the first time. Again analogously to the base case, Lemma 6.5 gives

$$\sum_{i\in I_k^{++}}[\mathbf{\Phi}]^+_{k,i} > \sum_{i=0}^{m-1}\left(1-\sum_{j=0}^{i}\frac{1}{M-j}\right)\overline{\Delta}$$

and we have an entry i satisfying

$$\boldsymbol{\gamma}_{k,i} = [\mathbf{\Phi}]^+_{k,i} > \frac{\sum_{i=0}^{m-1}\left(1-\sum_{j=0}^{i}\frac{1}{M-j}\right)}{M-m}\overline{\Delta}. \tag{6.2}$$

Regarding the sum, it is unclear what a negative numerator would mean here. Therefore, we distinguish the cases and show that a negative numerator is impossible.

Case A: $1-\sum_{j=0}^{m-1}\frac{1}{M-j} < 0$**.** We apply Lemma A.1 and obtain

$$\sum_{i\in I_k^{++}}\mathbf{\Phi}_{k,i} = \left((M-m)\sum_{j=0}^{m-1}\frac{1}{M-j}\right)\overline{\Delta} > (M-m)\overline{\Delta} \geq \left|I_k^{++}\right|\overline{\Delta}$$

because $m \leq \left|I_k^-\right| = M - \left|I_k^{++}\right|$. This situation only occurs if

$$\frac{1}{|I_k^{++}|}\sum_{i\in I_k^{++}}\boldsymbol{\gamma}_{k,i} = \frac{1}{|I_k^{++}|}\sum_{i\in I_k^{++}}\mathbf{\Phi}_{k,i} > \overline{\Delta}.$$

Consequently, we obtain $\boldsymbol{\gamma}_{k,i} > \overline{\Delta}$ for at least one $i \in I_k^{++}$. But this means $\boldsymbol{\gamma}_{k,i^*} \geq \overline{\Delta}$ when i^* denotes the rounding index and consequently,

$$\boldsymbol{\Phi}_{k,i^*} = \boldsymbol{\gamma}_{k,i^*} - h_k > \overline{\Delta} - h_k \geq 0.$$

We deduce that no negative entry was decreased when applying (SUR). Consequently, the sum of the negative entries cannot have broken the bound in this step, but this was the assumption and we have contradicted this case successfully.

Case B: $1 - \sum_{j=0}^{m-1} \frac{1}{M-j} \geq 0$**.** Again, we apply Lemma A.1 which yields

$$\boldsymbol{\gamma}_{k,i} = \boldsymbol{\Phi}_{k,i} > \sum_{j=0}^{m-1} \frac{1}{M-j} \overline{\Delta}$$

for $i \in I_k^{++}$ from (6.2). We denote the rounding index in cell k by i^* and note that $i^* \in J$ with $J \coloneqq \left\{ i_1^{k,-}, \ldots, i_m^{k,-} \right\} \subset I_k^-$. This holds true because only one entry can be reduced by (SUR) per cell. Plugging in the previous considerations, we obtain

$$\begin{aligned}
[\boldsymbol{\Phi}]_{k,i^*}^- &= h_k(1 - \boldsymbol{\alpha}_{k,i^*}) - \boldsymbol{\Phi}_{k-1,i^*} \\
&= h_k - \boldsymbol{\gamma}_{k,i^*} \\
&\leq \overline{\Delta} - \boldsymbol{\gamma}_{k,i^*} \\
&< \overline{\Delta} - \sum_{j=0}^{m-1} \frac{1}{M-j} \overline{\Delta}.
\end{aligned}$$

Using the initial assumption, we obtain

$$\sum_{i=0}^{m-1} \left(1 - \sum_{j=0}^{i} \frac{1}{M-j} \right) \overline{\Delta} < \sum_{i \in J \setminus \{i^*\}} [\boldsymbol{\Phi}]_{k,i}^- + \left(1 - \sum_{j=0}^{m-1} \frac{1}{M-j} \right) \overline{\Delta}$$

which gives

$$\sum_{i=0}^{m-2} \left(1 - \sum_{j=0}^{i} \frac{1}{M-j} \right) \overline{\Delta} < \sum_{i \in J \setminus \{i^*\}} [\boldsymbol{\Phi}]_{k,i}^- \leq \sum_{j=1}^{m-1} [\boldsymbol{\Phi}]_{k,i_j^{k,-}}^- = \left\| \boldsymbol{\psi}_{k,m-1}^- \right\|_1.$$

This contradicts the prerequisite (i.e. the induction hypothesis) because the left side is the bound from prerequisites and as this holds for the $m-1$ largest entries in $[\boldsymbol{\Phi}]_{k,\cdot}^-$, it also holds for all other subsets with $m-1$ elements in $[\boldsymbol{\Phi}]_{k,\cdot}^-$. Hence, the assumption that the claimed bound does not hold is false and the proof is complete. □

6.3.2 The bound on the sum-norm is tight

This bound is tight, which is demonstrated with the following algorithm. We use an equidistant rounding grid with $\overline{\Delta} = 1$.

Algorithm 6.1 Maximize $\| [\mathbf{\Phi}]^-_{k,\cdot} \|_1$

Input: $\mathbf{\Phi}_{0,\cdot} = 0$, $M > 1$, $t_{i+1} - t_i = 1$ for all $i \in \mathbb{N}$.

$k \leftarrow 0$

do

$k \leftarrow k + 1$

$\boldsymbol{\alpha}_{k,i} \leftarrow \begin{cases} \frac{1}{M-(k-1)} & \text{for } i \in I^+_{k-1} \\ 0 & \text{else} \end{cases}$

$\boldsymbol{\omega}_{k,\cdot} \leftarrow (\text{SUR})(\mathbf{\Phi}_{k-1,\cdot}, \boldsymbol{\alpha}_{k,\cdot}, 1)$

$\mathbf{\Phi}_{k,\cdot} \leftarrow \mathbf{\Phi}_{k-1,\cdot} + \boldsymbol{\alpha}_{k,\cdot} - \boldsymbol{\omega}_{k,\cdot}$

while $\frac{1}{M-k} + [\mathbf{\Phi}]^+_{k,i^{k,+}_1} < 1$

return $\boldsymbol{\alpha}_{1,\cdot}, \ldots, \boldsymbol{\alpha}_{k,\cdot}$

The algorithm terminates if a further iteration would make the positive entries of $\mathbf{\Phi}_{k,\cdot}$, all being equal to $\| [\mathbf{\Phi}]^+_{k,\cdot} \|_\infty$, exceed 1, which prevents a further increase in $\| [\mathbf{\Phi}]^-_{k,\cdot} \|_1$. To see this, we consider

$$\max_{j \in \{1,\ldots,M\}} \boldsymbol{\gamma}_{k,j} = \| [\mathbf{\Phi}]^+_{k,\cdot} \|_\infty,$$

and the fact that the increase in $\| [\mathbf{\Phi}]^-_{k,\cdot} \|_1$ decreases. The increase in iteration k is $1 - \max_{j \in \{1,\ldots,M\}} \boldsymbol{\gamma}_{k,j}$ and only happens if this quantity is positive, i.e. an entry changed from a positive to a negative value. Regarding the entries of $\boldsymbol{\alpha}_{k,\cdot}$ and $\boldsymbol{\omega}_{k,\cdot}$, we observe

$$\boldsymbol{\alpha}_{k,i} = \frac{1}{|I^+_{k-1}|}$$

for $i \in I^+_{k-1}$. Once an entry i^* was selected for rounding in iteration k_0, see the entries of $\boldsymbol{\omega}_{k,\cdot}$, we have $i^* \notin I^+_k$ for $k \geq k_0$ and $\boldsymbol{\alpha}_{k,i^*} = 0$ in later iterations $k > k_0$. These observations guide us towards proving the behavior of Algorithm 6.1.

Proposition 6.11. *Algorithm 6.1 terminates after a finite number of steps with iteration (grid cell) $k \in \mathbb{N}$ such that the constructed $\boldsymbol{\alpha}_{1,\cdot}, \ldots, \boldsymbol{\alpha}_{k,\cdot} \in \operatorname{conv} \mathbb{S}^M$ and the $\mathbf{\Phi}_{1,\cdot}, \ldots, \mathbf{\Phi}_{k,\cdot}$ induced by* (SUR) *on $h_1 = \ldots = h_k = \overline{\Delta}$ satisfy*

$$\left| I^-_k \right| = \min_{k \in \{1,\ldots,M-1\}} k \text{ s.t. } \sum_{i=0}^{k} \frac{1}{M-i} > 1$$

Furthermore,

$$\sum_{i=0}^{m-1} 1 - \sum_{j=0}^{i} \frac{1}{M-j} = \| [\mathbf{\Phi}]^-_{m,\cdot} \|_1$$

for all $m \leq \left| I^-_k \right|$.

Proof. In every iteration, Algorithm 6.1 produces the identity

$$\mathbf{\Phi}_{k,j} = \sum_{j=0}^{k-1} \frac{1}{M-(k-1)+j}$$

for all $j \in I_k^+$ inductively. The assignment $\boldsymbol{\alpha}_{k,\cdot}$ is well-defined if $|I_k^+| = M-k$. This holds true because we start with $|I_k^+| = M$ for $k=0$ and reduce I_k^+ by one entry in every one iteration as long as there are still entries left and $[\mathbf{\Phi}]_{k,j}^+ < 1$ for $j \in I_k^+$. When $[\mathbf{\Phi}]_{k+1,j}^+ > 1$ would happen in the next iteration, the termination criterion

$$[\mathbf{\Phi}]_{k,j}^+ + \frac{1}{M-k} > 1$$

is satisfied. For $k = M-1$, we have $[\mathbf{\Phi}]_{k,j}^+ + \frac{1}{M-(M-1)} = [\mathbf{\Phi}]_{k-1,j}^+ + 1 > 1$. Thus, the algorithm terminates after at most $M-1$ iterations and in all iterations there are positive entries left. Note that without loss of generality, we have $j = i_1^{k,+}$ as all positive entries assume the same value. For the negative part, we observe that if i^* is the rounding index in the k-th iteration, the corresponding integrated control deviation value is of

$$[\mathbf{\Phi}]_{k,i^*}^- = 1 - \sum_{j=0}^{k} \frac{1}{M-j}.$$

Summing over these entries along the iteration counter gives the claimed sum formula for the 1-norm as $\mathbf{\Phi}_{k,i^*}$ does not change in later iterations. The termination criterion is equivalent to $\sum_{i=0}^{k} \frac{1}{M-i} > 1$, which closes the proof as $k = |I_k^-|$ after every iteration and at termination. □

Remark 6.12. *Having established Proposition 6.8 and Proposition 6.11 helps to make some interesting observations. If $M \geq 3$ and m is large enough, it is possible that*

$$1 - \sum_{j=0}^{m-1} \frac{1}{M-j} < 0$$

as the sum is over a contiguous subset of the harmonic sequence. But these are the summands of the upper (!) bound of the $\| [\mathbf{\Phi}]_{k,\cdot}^- \|_1$. At first, this looks like a contradiction to what we have shown before. However, this simply tells us that the number of negative entries is traded off with $\| [\mathbf{\Phi}]_{k,\cdot}^- \|_1$ and one might not be able to maximize them simultaneously. The same holds true for the strictly negative indices as a similar chain of arguments can be carried out in this case. The example below shows such a situation.

Example 6.13. *We set $\overline{\Delta} = 1$, $M = 5$, $N = 4$ and use the values of $\boldsymbol{\alpha}_{\cdot,\cdot}$ given in the even columns of Table 6.1. Applying* (SUR) *yields the values of $\mathbf{\Phi}$ given in the odd columns of Table 6.1. For $I_4^{--} = 4$, we obtain the bound*

$$\| [\mathbf{\Phi}]_{4,\cdot}^- \|_1 \leq \frac{1}{5} + \frac{1}{4} + \frac{1}{3} + \frac{1}{2}$$

from Proposition 6.8 where the last summand of the formula in Proposition 6.8 is $1 - \left(\frac{1}{5} + \frac{1}{4} + \frac{1}{3} + \frac{1}{2}\right) < 0$. *Consequently, although we have more strictly negative indices in* $\mathbf{\Phi}_{4,\cdot}$ *than in* $\mathbf{\Phi}_{3,\cdot}$*, the bound is tighter as the one we have in the third iteration*

$$\| [\mathbf{\Phi}]^-_{3,\cdot} \|_1 \leq 2\left(\frac{1}{5} + \frac{1}{4} + \frac{1}{3}\right).$$

We make a check and see

$$\| [\mathbf{\Phi}]^-_{3,\cdot} \|_1 = 2\left(\frac{1}{5} + \frac{1}{4} + \frac{1}{3}\right),$$

i.e. the bound is reached in the third iteration and

$$\| [\mathbf{\Phi}]^-_{4,\cdot} \|_1 = \frac{59}{60} < \frac{77}{60} = \frac{1}{5} + \frac{1}{4} + \frac{1}{3} + \frac{1}{2},$$

i.e. the bound holds but is not reached in the fourth iteration. Thus, we have validated the results from Proposition 6.8 and observe that in order to get the fourth strictly negative entry in $\mathbf{\Phi}_{4,\cdot}$*, we had to sacrifice some fraction of* $\| [\mathbf{\Phi}]^-_{3,\cdot} \|_1$.

Table 6.1: Input $\boldsymbol{\alpha}_{\cdot,\cdot}$ and the resulting values of $\mathbf{\Phi}_{\cdot,\cdot}$ for the application of (SUR).

i	$\boldsymbol{\alpha}_{1,i}$	$\mathbf{\Phi}_{1,i}$	$\boldsymbol{\alpha}_{2,i}$	$\mathbf{\Phi}_{2,i}$	$\boldsymbol{\alpha}_{3,i}$	$\mathbf{\Phi}_{3,i}$	$\boldsymbol{\alpha}_{4,i}$	$\mathbf{\Phi}_{4,i}$
1	$\frac{1}{5}$	$\frac{1}{5}$	$\frac{1}{4}$	$\frac{9}{20}$	$\frac{1}{3}$	$\frac{47}{60}$	$\frac{1}{5}$	$\frac{59}{60}$
2	$\frac{1}{5}$	$\frac{1}{5}$	$\frac{1}{4}$	$\frac{9}{20}$	$\frac{1}{3}$	$\frac{47}{60}$	$\frac{1}{5}$	$-\frac{1}{60}$
3	$\frac{1}{5}$	$\frac{1}{5}$	$\frac{1}{4}$	$\frac{9}{20}$	$\frac{1}{3}$	$-\frac{13}{60}$	$\frac{1}{5}$	$-\frac{1}{60}$
4	$\frac{1}{5}$	$\frac{1}{5}$	$\frac{1}{4}$	$-\frac{11}{20}$	0	$-\frac{11}{20}$	$\frac{1}{5}$	$-\frac{21}{60}$
5	$\frac{1}{5}$	$-\frac{4}{5}$	0	$-\frac{4}{5}$	0	$-\frac{4}{5}$	$\frac{1}{5}$	$-\frac{36}{60}$

6.3.3 Maximizing the max-norm of the integrated control deviation

Having established the bounds on $\| [\mathbf{\Phi}]^-_{k,\cdot} \|_1$, we need another ingredient to prove the desired bound on $\| [\mathbf{\Phi}]^+_{k,\cdot} \|_\infty$. It is stated in the following proposition.

Proposition 6.14. *Let* $k \in \mathbb{N}$ *such that* $|I^+_k| \geq m$ *for some* $m \in \{1, \ldots, M\}$. *All* $\boldsymbol{\alpha}_{1,\cdot}, \ldots, \boldsymbol{\alpha}_{k,\cdot} \in \operatorname{conv} \mathbb{S}^M$ *can be extended to* $\boldsymbol{\alpha} \in \mathcal{A}$ *with* $\boldsymbol{\alpha}_{k+1,\cdot}, \ldots, \boldsymbol{\alpha}_{k+\ell,\cdot}$ *such that*

$$\left\|\boldsymbol{\psi}^+_{k,m}\right\|_1 = \left\|\boldsymbol{\psi}^+_{k+\ell,m}\right\|_1$$

and

$$[\mathbf{\Phi}]^+_{k+\ell,i_1^{k+\ell,+}} = \ldots = [\mathbf{\Phi}]^+_{k+\ell,i_m^{k+\ell,+}} = \frac{1}{m}\left\|\boldsymbol{\psi}^+_{k,m}\right\|_1$$

for some $\ell \in \mathbb{N}$.

Proof. The proof is a direct consequence of the following Proposition 6.18. □

In the following, we aim to confine the quantity S_1, which is defined as

$$S_1 \coloneqq \sup_{\boldsymbol{\alpha}\in\mathcal{A}} \sup_{k\in\mathbb{N}} \left\|\boldsymbol{\psi}^+_{k,1}\right\|_1 = \sup_{\boldsymbol{\alpha}\in\mathcal{A}} \sup_{k\in\mathbb{N}} [\boldsymbol{\Phi}]^+_{k,i_1^{k,+}} = \sup_{\boldsymbol{\alpha}\in\mathcal{A}} \sup_{k\in\mathbb{N}} \left\|[\boldsymbol{\Phi}]^+_{k,\cdot}\right\|_\infty .$$

Note that S_1 has to be finite byProposition 6.8 and Lemma 6.5. Let us assume that S_1 is attained by the sequence of $(\boldsymbol{\Phi}_{k,\cdot})_k$ produced by (SUR) from the sequence $(\boldsymbol{\alpha}_{k,\cdot})_k$. The case that S_1 is not attained will be handled later by introducing sufficient epsilons into the arguments. We start with a lemma on the relationship between the two largest positive entries, which we find quite instructive. It uses the same ideas as the subsequent proofs but shows the most easy case.

Lemma 6.15. *Let $\boldsymbol{\alpha} \in \mathcal{A}$ such that $[\boldsymbol{\Phi}]^+_{k,i_1^{k,+}} = S_1$ for some $k \in \mathbb{N}$. If $|I_k^+| \geq 2$, we have*

$$[\boldsymbol{\Phi}]^+_{k,i_2^{k,+}} = S_1 - \overline{\Delta}.$$

Proof. We show that the options $[\boldsymbol{\Phi}]^+_{k,i_2^{k,+}} > S_1 - \overline{\Delta}$ and $[\boldsymbol{\Phi}]^+_{k,i_2^{k,+}} < S_1 - \overline{\Delta}$ lead to contradictions.

First, we assume $[\boldsymbol{\Phi}]^+_{k,i_2^{k,+}} < S_1 - \overline{\Delta}$. Then, there exists a minimal $k_0 \in \mathbb{N}$ such that $[\boldsymbol{\Phi}]^+_{k_0,i_1^{k_0,+}} = S_1$ and $[\boldsymbol{\Phi}]^+_{k_0,i_2^{k_0,+}} < S_1 - \overline{\Delta}$. If $i_1^{k_0,+}$ was the rounding index in cell k_0, we would have

$$S_1 \geq [\boldsymbol{\Phi}]^+_{k_0-1,i_1^{k_0,+}} \geq [\boldsymbol{\Phi}]^+_{k_0,i_1^{k_0,+}} = S_1$$

and

$$[\boldsymbol{\Phi}]^+_{k_0-1,i_2^{k_0,+}} \leq [\boldsymbol{\Phi}]^+_{k_0,i_2^{k_0,+}} < S_1 - \overline{\Delta},$$

which is a contradiction because k_0 is the first (minimal) cell, in which this situation occurs. If $i^* \neq i_1^{k_0,+}$ was the rounding index in cell k_0, we would have

$$[\boldsymbol{\Phi}]^+_{k_0,i^*} + h_{k_0} = \boldsymbol{\gamma}_{k_0,i^*} \geq \boldsymbol{\gamma}_{k_0,i_1^{k_0,+}} = [\boldsymbol{\Phi}]^+_{k_0,i_1^{k_0,+}} = S_1,$$

which gives $[\boldsymbol{\Phi}]^+_{k_0,i^*} \geq S_1 - \overline{\Delta}$ and contradicts $[\boldsymbol{\Phi}]^+_{k,i_2^{k,+}} < S_1 - \overline{\Delta}$.

Second, we handle the case $x \coloneqq [\boldsymbol{\Phi}]^+_{i_2^{k,+}} > S_1 - \overline{\Delta}$. Then, Proposition 6.14 gives us $\boldsymbol{\alpha} \in \mathcal{A}$ and an iterate k_1 such that

$$[\boldsymbol{\Phi}]^+_{k_1,i_1^{k_1,+}} = [\boldsymbol{\Phi}]^+_{k_1,i_2^{k_1,+}} = \frac{S_1 + x}{2}.$$

Due to the padding of the grid cells with $\lambda(\mathcal{T}_k) = \overline{\Delta}$ to infinity, we can assume that there exists $\ell \in \mathbb{N}$ such that $h_{k_1+\ell} = \overline{\Delta}$. We can assume $\boldsymbol{\alpha}_{k,i_1^{k_1,+}} = 1$ for $k_1 \leq k \leq k_1 + \ell - 1$ and $\boldsymbol{\alpha}_{k_1+\ell,i_1^{k_1,+}} = \boldsymbol{\alpha}_{k_1+\ell,i_1^{k_1,+}} = \frac{1}{2}$. Then, the application of (SUR) yields

$$[\boldsymbol{\Phi}]^+_{k_1+\ell,i_1^{k_1+\ell,+}} = \frac{S_1 + x}{2} + \frac{\overline{\Delta}}{2} > S_1,$$

which contradicts the definition of S_1. □

The following lemma establishes a bound on the supremum of the sum of the j largest positive entries, denoted by S_j, depending on the supremum of the sum of all positive (nonnegative) entries, denoted by S.

Lemma 6.16. *Let* $S \coloneqq \sup_{\boldsymbol{\alpha}\in\mathcal{A}} \sup_{k\in\mathbb{N}} \left\| [\boldsymbol{\Phi}]^+_{k,\cdot} \right\|_1$ *and* $m \in \mathbb{N}$ *be such that* $(m-1)\overline{\Delta} < S \leq m\overline{\Delta}$. *Then,*

$$\begin{aligned}\sup_{\boldsymbol{\alpha}\in\mathcal{A}} \sup_{k\in\mathbb{N}} \left\|\boldsymbol{\psi}^+_{k,j}\right\|_1 &= j\left(\frac{S}{m\overline{\Delta}} + \frac{1}{m} + \ldots + \frac{1}{j+1}\right)\overline{\Delta} \\ &= j\left(\frac{S}{m\overline{\Delta}} + \sum_{i=j+1}^{m} \frac{1}{i}\right)\overline{\Delta} \\ &=: S_j\end{aligned}$$

for all $j \in \{1,\ldots,m\}$.

Proof. The proof employs backward induction to prove the claim. We start with the base case and reduce to the setting $|I^+_k| = m$. Let $S - \| [\boldsymbol{\Phi}]^+_{k,\cdot} \|_1 < \varepsilon$, but $|I^{++}_k| = m + p$ for some $p \geq 1$. Then, $[\boldsymbol{\Phi}]^+_{k,i} < \overline{\Delta}$ for $i \in I^{++}_k$ by Proposition 6.14 without loss of generality. Indeed, for one $i \in I^{++}_k$, we set $\boldsymbol{\alpha}_{k+1,i}$ such that $\boldsymbol{\gamma}_{k+1,i} = \overline{\Delta}$ and distribute $h_{k+1} - \boldsymbol{\alpha}_{k+1,i}h_{k+1}$ evenly among the remaining indices in $I^{++}_k\backslash\{i\}$. We can achieve $\boldsymbol{\gamma}_{k+1,i} = \overline{\Delta}$ because we may assume $h_{k+1} = \overline{\Delta}$ as in the proof of Lemma 6.15. Then, the application of (SUR) gives $[\boldsymbol{\Phi}]^+_{k+1,i} = 0$ and $|I^{++}_{k+1}| = m + p - 1$. This can be repeated until $|I^{++}_k| = m$. Thus, we obtain

$$S_m = m\frac{S}{m}\frac{\overline{\Delta}}{\overline{\Delta}} = S = \sup_{\boldsymbol{\alpha}\in\mathcal{A}} \sup_{k\in\mathbb{N}} \|\boldsymbol{\psi}^+_{k,m}\|_1,$$

which establishes the base case. Let us assume

$$\sup_{\boldsymbol{\alpha}\in\mathcal{A}} \sup_{k\in\mathbb{N}} \left\|\boldsymbol{\psi}^+_{k,j}\right\|_1 = S_j$$

for some $j \leq k$. We show the upper and lower limits separately to complete the induction $j \to j-1$.

Inequality $\sup_{\boldsymbol{\alpha}\in\mathcal{A}} \sup_{k\in\mathbb{N}} \left\|\boldsymbol{\psi}^+_{k,j-1}\right\|_1 \geq S_{j-1}$**:** Let $\varepsilon > 0$ and $k \in \mathbb{N}$ such that

$$\left\|\boldsymbol{\psi}^+_{k,j}\right\|_1 = S_j - \varepsilon \text{ and } \left\|\boldsymbol{\psi}^+_{k,j}\right\|_1 - \varepsilon > m - 1$$

for some $\boldsymbol{\alpha} \in \mathcal{A}$. We apply Proposition 6.14 and obtain a possible continuation of $\boldsymbol{\alpha}_{1\leq\ell\leq k,\cdot}$ with $k_0 \in \mathbb{N}$, $\boldsymbol{\alpha}_{k+1,\cdot},\ldots,\boldsymbol{\alpha}_{k_0,\cdot}$ such that

$$\left\|\boldsymbol{\psi}^+_{k,j}\right\|_1 = \left\|\boldsymbol{\psi}^+_{k_0,j}\right\|_1$$

and

$$[\boldsymbol{\Phi}]^+_{k_0,i_1^{k_0,+}} = \ldots = [\boldsymbol{\Phi}]^+_{k_0,i_j^{k_0,+}} = \frac{S_j - \varepsilon}{j}.$$

We construct

$$\boldsymbol{\alpha}_{k_0+1,i} \coloneqq \begin{cases} \frac{1}{j} & : \quad i \in \left\{ i_1^{k_0,+}, \ldots, i_j^{k_0,+} \right\} \\ 0 & : \quad \text{else} \end{cases}$$

which yields $\boldsymbol{\gamma}_{k_0+1,i_1^{k_0,+}} = \ldots = \boldsymbol{\gamma}_{k_0+1,i_j^{k_0,+}}$. With the same consideration as above, we can assume $h_{k_0+1} = \overline{\Delta}$ as we want to find a lower bound for a $\sup$. Then,

$$[\boldsymbol{\Phi}]^+_{k_0+1,i_1^{k_0+1,+}} = \ldots = [\boldsymbol{\Phi}]^+_{k_0+1,i_{j-1}^{k_0+1,+}} = \frac{S_j - \varepsilon}{j} + \frac{1}{j}\overline{\Delta}$$

for $j-1$ entries and one of them was reduced as the largest entries of $[\boldsymbol{\Phi}]^+_{k_0,\cdot}$ were considered when setting $\boldsymbol{\alpha}_{k_0+1,i}$. Consequently,

$$\begin{aligned} \left\| \boldsymbol{\psi}^+_{k_0+1,j-1} \right\|_1 &= (j-1)\left(\frac{S_j}{j} + \frac{1}{j} - \frac{\varepsilon}{j\overline{\Delta}} \right)\overline{\Delta} \\ &= (j-1)\left(\frac{S}{m\overline{\Delta}} + \frac{1}{m} + \ldots + \frac{1}{j+1} + \frac{1}{j} \right)\overline{\Delta} - \frac{j-1}{j}\varepsilon = S_{j-1} - \frac{j-1}{j}\varepsilon. \end{aligned}$$

Driving $\varepsilon \to 0$, we obtain

$$\sup_{\boldsymbol{\alpha}\in\mathcal{A}} \sup_{k\in\mathbb{N}} \left\| \boldsymbol{\psi}^+_{k,j-1} \right\|_1 \geq S_{j-1}.$$

Inequality $\sup_{\boldsymbol{\alpha}\in\mathcal{A}} \sup_{k\in\mathbb{N}} \left\| \boldsymbol{\psi}^+_{k,j-1} \right\|_1 \leq S_{j-1}$: We proceed by contradiction and assume $\sup_{k\in\mathbb{N}} \|\boldsymbol{\psi}^+_{k,j-1}\|_1 > S_{j-1}$. Then, there exists a minimal $k_0 \in \mathbb{N}$ such that $\|\boldsymbol{\psi}^+_{k_0,j-1}\|_1 > S_{j-1}$, i.e. ,

$$\left\| \boldsymbol{\psi}^+_{k_0,j-1} \right\|_1 > (j-1)\left(\frac{S}{m\overline{\Delta}} + \frac{1}{m} + \ldots + \frac{1}{j} \right)\overline{\Delta}. \tag{6.3}$$

By definition of (SUR), the rounding index i^* satisfies $i^* \notin \left\{ i_1^{k_0,+}, \ldots, i_{j-1}^{k_0,+} \right\}$ and

$$[\boldsymbol{\Phi}]^+_{k_0,i^*} > \left(\frac{S}{m\overline{\Delta}} + \frac{1}{m} + \ldots + \frac{1}{j} \right)\overline{\Delta} - h_{k_0} \tag{6.4}$$

because $\boldsymbol{\gamma}_{k_0,i^*}$ is a maximizing entry of $\boldsymbol{\gamma}_{k_0,\cdot}$ and at least one of the $j-1$ largest entries has to be greater or equal than the mean of the $j-1$ largest entries. Otherwise, no increase (to the state $\|\boldsymbol{\psi}^+_{k_0,j-1}\|_1 > S_j$) would have been possible. By definition of $\boldsymbol{\psi}^+_{k_0,j}$ we observe

$$\left\| \boldsymbol{\psi}^+_{k_0,j} \right\|_1 \geq \left\| \boldsymbol{\psi}^+_{k_0,j-1} \right\|_1 + [\boldsymbol{\Phi}]^+_{k_0,i^*}.$$

We plug in (6.3) and (6.4) and obtain

$$\begin{aligned} \left\| \boldsymbol{\psi}^+_{k_0,j} \right\|_1 &> j\left(\frac{S}{m\overline{\Delta}} + \frac{1}{m} + \ldots + \frac{1}{j} \right)\overline{\Delta} - h_k \\ &\geq j\left(\frac{S}{m\overline{\Delta}} + \frac{1}{m} + \ldots + \frac{1}{j+1} \right)\overline{\Delta}, \end{aligned}$$

which contradicts the induction hypothesis, which in turn proves the desired bound. □

We have established a bound on the sum-norm in Lemma 6.16, which is key to prove our ultimate claim on $\sup_{k\in\mathbb{N}} \|\mathbf{\Phi}_{k,\cdot}\|_\infty$.

Theorem 6.17 (Theorem 6.7)**.** *Consider* (SUR) *applied to a relaxed control $\boldsymbol{\alpha}$ on some rounding grid. Let $\mathbf{\Phi}_{k,\cdot}$ be the resulting integrated control deviation at cell k. Then,*

$$\sup_{1\leq k\leq N} \|\mathbf{\Phi}_{k,\cdot}\|_\infty \leq \overline{\Delta} \sum_{i=2}^{M} \frac{1}{i}.$$

Proof. For $M = 2$ and $M = 3$, we have $\| [\mathbf{\Phi}]^-_{k,\cdot} \|_1 < \overline{\Delta}$ and thus, by virtue of Lemmas 6.5 and A.1 and Proposition 6.8, $\| [\mathbf{\Phi}]^+_{k,\cdot} \|_\infty \leq \| [\mathbf{\Phi}]^-_{k,\cdot} \|_1 \leq \sum_{j=0}^{M-2} \frac{1}{M-j} \overline{\Delta}$.

Let $M \geq 4$. Then, the optimization problem

$$\min_{n\in\{1,\dots,M-1\}} n \text{ s.t. } \sum_{i=0}^{n} \frac{1}{M-i} > 1$$

admits a minimizer n^*, which gives the value of $|I^-_k|$ to extremize the sum-norm in Proposition 6.8. For $S \coloneqq \sup_{\boldsymbol{\alpha}\in\mathcal{A}} \sup_{k\in\mathbb{N}} \| [\mathbf{\Phi}]^+_{k,\cdot} \|_1$, we obtain

$$S = \left(\sum_{i=0}^{n^*-1} 1 - \sum_{j=0}^{i} \frac{1}{M-j} \right) \overline{\Delta} = (M-n^*) \sum_{i=0}^{n^*-1} \frac{1}{M-i} \overline{\Delta}.$$

Here, the first equality follows from the tightness of the bound established in Proposition 6.11, Lemma 6.5 and Proposition 6.11, which produces the tightness as Algorithm 6.1 can be assumed to begin after N cells with $\boldsymbol{\alpha}_{1\leq k\leq N,\cdot} \equiv e_1$ and we may continue the sequence $(h_k)_k$ with $h_k = \overline{\Delta}$ for $k > N$. The second equality is due to Lemma A.1. We define m as in Lemma 6.16, which gives

$$m = \left\lceil \frac{S}{\overline{\Delta}} \right\rceil = \left\lceil (M-n^*) \sum_{i=0}^{n^*-1} \frac{1}{M-i} \right\rceil.$$

From the construction of n^*, we have $\sum_{i=0}^{n^*-1} \frac{1}{M-i} \leq 1$, which gives $m \leq M - n^*$. Assume $m < M - n^*$. This implies

$$(M-n^*) \sum_{i=0}^{n^*-1} \frac{1}{M-i} \leq M - n^* - 1 \Leftrightarrow \sum_{i=0}^{n^*-1} \frac{1}{M-i} \leq \frac{M-n^*-1}{M-n^*},$$

but this means $\sum_{i=0}^{n^*} \frac{1}{M-i} \leq 1$, which contradicts the construction of n^*. Thus, we have $m = M - n^*$. We insert the considerations into Lemma 6.16, which gives

$$\begin{aligned} \frac{S_1}{\overline{\Delta}} &= \frac{(M-n^*) \sum_{i=0}^{n^*-1} \frac{1}{M-i}}{m} + \frac{1}{m} + \dots + \frac{1}{2} \\ &= \sum_{i=0}^{n^*-1} \frac{1}{M-i} + \frac{1}{M-n^*} + \dots + \frac{1}{2} = \sum_{i=2}^{M} \frac{1}{i}. \end{aligned}$$

The second equality follows from the identity $m = M - n^*$. This proves the claim. □

Finally, we prove Proposition 6.14. We exploit the fact that $\mathbf{\Phi}_{k,\cdot}$ only depends on $\boldsymbol{\alpha}_{1,\cdot}, \dots, \boldsymbol{\alpha}_{k,\cdot}$ or, suited to our case, $\mathbf{\Phi}_{k+\ell,\cdot}$ only depends on $\mathbf{\Phi}_{k,\cdot}$ and $\boldsymbol{\alpha}_{k+1,\cdot}, \dots, \boldsymbol{\alpha}_{k+\ell,\cdot}$. Thus, we state an algorithm, Algorithm 6.2, and prove that it generates the desired $\boldsymbol{\alpha}_{k+1,\cdot}, \dots, \boldsymbol{\alpha}_{k+\ell,\cdot}$ for every starting configuration $\mathbf{\Phi}_{k,\cdot}$.

Algorithm 6.2 Equilibrate high entries

Input: $\mathbf{\Phi}_{k,\cdot}$, $j \in \{1, \dots, M\}$ such that $[\mathbf{\Phi}]^+_{k,i_j^{k,+}} > 0$.

$\bar{\mathbf{\Phi}} \leftarrow \frac{1}{j}\|\boldsymbol{\psi}^+_{k,j}\|_1$
$i \leftarrow 0$
while $\exists \ell \in \left\{i_1^{k,+}, \dots, i_j^{k,+}\right\} : [\mathbf{\Phi}]^+_{k+i,\ell} \neq \bar{\mathbf{\Phi}}$ **do**
  if $h_{k+i+1} \neq \overline{\Delta}$ **then**
    $$\boldsymbol{\alpha}_{k+i+1,\ell} \leftarrow \begin{cases} 1 & \text{if } \ell = i_1^{k+i} \\ 0 & \text{else} \end{cases}$$
  else
    $\underline{\ell} \leftarrow \arg\min \left\{\mathbf{\Phi}_{k+i,\ell} : \ell \in \{i_1^{k,+}, \dots, i_j^{k,+}\}\right\}$
    $\overline{\ell} \leftarrow \arg\max \left\{\mathbf{\Phi}_{k+i,\ell} : \ell \in \{i_1^{k,+}, \dots, i_j^{k,+}\}\right\}$
    $$\boldsymbol{\alpha}_{k+i+1,\ell} \leftarrow \begin{cases} \min\left\{\frac{\bar{\mathbf{\Phi}} - [\mathbf{\Phi}]^+_{k+i,\underline{\ell}}}{h_{k+i+1}}, 1.\right\} & \ell = \underline{\ell} \\ 1 - \boldsymbol{\alpha}_{k+i+1,\underline{\ell}} & \ell = \overline{\ell} \\ 0 & \text{else} \end{cases}$$
  end if
  $\boldsymbol{\omega}_{k+i+1,\cdot} \leftarrow (\text{SUR})(\mathbf{\Phi}_{k+i,\cdot}, \boldsymbol{\alpha}_{k+i+1,\cdot}, h_{k+i+1})$.
  $\mathbf{\Phi}_{k+i+1,\cdot} \leftarrow \mathbf{\Phi}_{k+i,\cdot} + \boldsymbol{\alpha}_{k+i+1,\cdot} - \boldsymbol{\omega}_{k+i+1,\cdot}$
  $i \leftarrow i + 1$
end while
return $\boldsymbol{\alpha}_{k+1,\cdot}, \dots, \boldsymbol{\alpha}_{k+i,\cdot}$

Proposition 6.18. *Algorithm 6.2 terminates after a finite number i of steps with $\boldsymbol{\alpha}_{k+1,\cdot}, \dots, \boldsymbol{\alpha}_{k+i,\cdot}$ such that*

$$\frac{1}{j}\left\|\boldsymbol{\psi}^+_{k,j}\right\|_1 = [\mathbf{\Phi}]^+_{k+i,i_1^k} = \dots = [\mathbf{\Phi}]^+_{k+i,i_j^k}.$$

Proof. Due to the termination criterion of Algorithm 6.2, it terminates correctly if it terminates. It remains to show that it terminates after finitely many steps. Furthermore, the algorithm only induces a change from $\mathbf{\Phi}_{k+i,\cdot}$ to $\mathbf{\Phi}_{k+i+1,\cdot}$ if $h_{k+i} = \overline{\Delta}$. As the sequence of grid cells is continued with grid cells of volume $\overline{\Delta}$ for $k > N$, the algorithm cannot get stuck in the if-branch and we can restrict considerations to the else-branch.

One of the smallest elements of the j largest entries of $[\mathbf{\Phi}]^+$ is increased from one iteration to the next and one of the largest is decreased. Consider an iteration i: the entry $\underline{\ell}$ is strictly less and the entry $\overline{\ell}$ is strictly greater than $\bar{\mathbf{\Phi}}$ before the iteration

because otherwise, the algorithm would have terminated. Furthermore, the assignment

$$\boldsymbol{\alpha}_{k+i+1,\underline{\ell}} = \min\left\{\frac{\bar{\boldsymbol{\Phi}} - [\boldsymbol{\Phi}]^+_{k+i,\underline{j}}}{h_{k+i+1}}, 1.\right\}$$

ensures

$$\boldsymbol{\gamma}_{k+i+1,\underline{\ell}} \leq \bar{\boldsymbol{\Phi}} < \boldsymbol{\gamma}_{k+i,\overline{\ell}}$$

yielding that the entry $\underline{\ell}$ is increased and the entry $\overline{\ell}$ (or another maximizing entry) is decreased after the application of (SUR). All other entries remain unchanged. Thus, the sum of the j largest entries under consideration does not change during Algorithm 6.2.

Consider an iteration $k+i$ and assume

$$[\boldsymbol{\Phi}]^+_{k+i,\underline{j}} + h_{k+i+1} \geq \bar{\boldsymbol{\Phi}}.$$

Then, the number of entries with $[\boldsymbol{\Phi}]^+_{k+i+1,\ell} \neq \bar{\boldsymbol{\Phi}}$, $\ell \in \{i^k_1, \dots, i^k_j\}$ is reduced by one and the entry is never selected again by $\arg\min$ and $\arg\max$. Assume $\boldsymbol{\Phi}_{k,\underline{j}} + h_{k+i+1} < \bar{\boldsymbol{\Phi}}$. Then,

$$\sum_{\ell=1}^{j} \left| [\boldsymbol{\Phi}]^+_{k+i+1,i^k_\ell} - \bar{\boldsymbol{\Phi}} \right| < \sum_{\ell=1}^{j} \left| [\boldsymbol{\Phi}]^+_{k+i,i^k_j} - \bar{\boldsymbol{\Phi}} \right| - \frac{1}{j-1} h_{k+i+1}$$

because at most $j-1$ entries can be of greater value than $\bar{\boldsymbol{\Phi}}$ and as $h_{k+i+1} \geq \inf_{k\in\mathbb{N}} h_k = \min_{k\in\{1,\dots,N\}} h_k > 0$, we have a constant reduction of this norm and terminate after a finite number of steps or the number of indices violating the termination criterion is reduced by one. Repeating this argument at most $j-1$ times (one entry is always greater than $\bar{\boldsymbol{\Phi}}$ as long as there is no termination), we get termination because all entries under consideration equal $\bar{\boldsymbol{\Phi}}$. □

6.4 A bound on the integrality gap for (SUR-VC)

Some of the results on the integrality gap that are proved below hold for different restrictions on the sets of admissible rounding indices $(F_k)_{k\in\mathbb{N}}$ in (SUR-GEN). Therefore, we introduce an assumption that holds in particular for (SUR-VC), but is slightly more general, e.g. it also holds for the standard case (SUR).

Assumption 6.19 (Admissible indices for rounding, [82, Ass. 4.1])**.** *For all $k \in \mathbb{N}$, we have*

$$\{1 \leq i \leq M : \boldsymbol{\alpha}_{k,j} > 0\} \subset F_k$$

during the execution of (SUR-GEN)*.*

Theorem 6.20 ([82, Thm 4.8])**.** *Let $\boldsymbol{\omega}$ be a binary control computed with* (SUR-GEN)*, satisfying Assumption 6.19, from a relaxed control $\boldsymbol{\alpha}$ on a rounding grid with mesh size $\overline{\Delta}$. Then,*

$$d(\boldsymbol{\omega}, \boldsymbol{\alpha}) \leq (M+1)\overline{\Delta}.$$

Outline of proof. Again, we employ Proposition 6.4 and prove the bound by verifying

$$\sup_{k\in\mathbb{N}} \|\mathbf{\Phi}_{k,\cdot}\|_\infty \leq (M+1)\overline{\Delta} \tag{6.5}$$

for all $\boldsymbol{\alpha} \in \mathcal{A}$ and $(h_k)_k \subset (0, \overline{\Delta}]$ with $h_k = \overline{\Delta}$ for $k \geq N$. For (6.5) as well as many of the following considerations, it is important to bear in mind that the $\mathbf{\Phi}_{k,i}$ depend on $\boldsymbol{\alpha}$. We will frequently exploit the fact that the $\sup$ has to hold for a $\mathbf{\Phi}$ arising from a specific $\boldsymbol{\alpha}$. More precisely, we will take advantage of the fact that $\mathbf{\Phi}_{k,\cdot}$ depends only on $\boldsymbol{\alpha}_{1\leq\ell\leq k,\cdot}$. This allows us to discuss $\mathbf{\Phi}_{k,\cdot}$ independent of or for all $\boldsymbol{\alpha}_{k<\ell<\infty,\cdot}$. We give preparatory lemmata, which establish

1. an upper bound on the entries of $\mathbf{\Phi}_{k,\cdot}$ that were admissible for rounding, but not chosen by (SUR-GEN);
2. lower bounds on entries of $\mathbf{\Phi}_{k,\cdot}$ that have the same sign if we can bind the sum of ℓ of them from above and the sum of $\ell+1$ from below;
3. upper bounds on sums of a set of positive / negative entries of $\mathbf{\Phi}_{k,\cdot}$ depending on the sum of the largest positive / negative entries in $\mathbf{\Phi}_{k-1,\cdot}$.

These preparations are used to prove an induction from an assumed break of the bound for some $\boldsymbol{\alpha} \in \mathcal{A}$, which yields the existence of a corresponding excession of the 1-norm, which would imply that all entries of $\mathbf{\Phi}_{k,\cdot}$ are strictly positive (negative), which contradicts the elementary properties of (SUR-GEN) given in Lemma 6.5.

Proving Theorem 6.20. The first preparatory lemma follows directly from the prerequisites of Theorem 6.20 and Definition 5.3. The following lemmata hold for a fixed $\boldsymbol{\alpha} \in \mathcal{A}$.

Lemma 6.21 ([82, Lem. 4.3])**.** *Let $k \in \mathbb{N}$, $\mathbf{\Phi}_{k,i^*} < \mathbf{\Phi}_{k-1,i^*}$ and let i^* be the rounding index in the k-th cell. Then, for all $j \in F_k \backslash \{i^*\}$*

$$\mathbf{\Phi}_{k,j} - \mathbf{\Phi}_{k,i^*} \leq h_k.$$

Furthermore, if $F_k = \{i^*\}$ and Assumption 6.19 holds, we get

$$\mathbf{\Phi}_{k,\cdot} = \mathbf{\Phi}_{k-1,\cdot}$$

Proof. We assume converse of the first claim. Then,

$$\begin{aligned}\mathbf{\Phi}_{k-1,j} + \boldsymbol{\alpha}_{k,j} h_k &= \mathbf{\Phi}_{k,j} \\ &> \mathbf{\Phi}_{k,i^*} + h_k \\ &= \mathbf{\Phi}_{k-1,i^*} + \boldsymbol{\alpha}_{k,i^*} h_k - h_k + h_k = \mathbf{\Phi}_{k-1,i^*} + \boldsymbol{\alpha}_{k,i^*} h_k,\end{aligned}$$

which contradicts i^* being the rounding index. If $F_k = \{i^*\}$, Lemma 6.5 and Assumption 6.19 yield $\boldsymbol{\alpha}_{k,i^*} = 1$ and establish the second claim. □

The next lemma gives a lower bound on entries of $[\mathbf{\Phi}]^{\pm}_{k,\cdot}$, i.e. on the infinity norm for certain subsets of the entries, from bounds on $\|\boldsymbol{\psi}^J_{k,\ell}\|_1$ for some $J \subset I^{\pm}_k$.

Lemma 6.22 ([82, Lem. 4.6]). *Let $\mathfrak{s} \in \{+,-\}$, $k \in \mathbb{N}$, $\ell < |J|$, $J \subset I^{\mathfrak{s}}_k$ and let the following bounds hold for some $\xi \geq \ell\overline{\Delta}$ and $\zeta \geq 0$:*

$$1. \quad \left\|\boldsymbol{\psi}^J_{k,\ell}\right\|_1 \leq \sum_{m=1}^{\ell} \xi - (m-1)\overline{\Delta}, \tag{6.6}$$

$$2. \ \left\|\boldsymbol{\psi}^J_{k,\ell+1}\right\|_1 > \left(\sum_{m=1}^{\ell+1} \xi - (m-1)\overline{\Delta}\right) - \zeta. \tag{6.7}$$

Then, the bound

$$\left|\mathbf{\Phi}_{k,j^k_m}\right| > \xi - \ell\overline{\Delta} - \zeta \underset{\text{if } \xi \geq (\ell+1)\overline{\Delta}, \zeta = 0}{\geq} \overline{\Delta}$$

holds for $1 \leq m \leq \ell+1$.

Proof. Due to the order encoded by the j^k_m, the claim holds if it holds for $m = \ell+1$. We assume the converse, i.e. $|\mathbf{\Phi}_{k,j^k_{\ell+1}}| \leq \xi - \ell\overline{\Delta} - \zeta$. Then,

$$\begin{aligned}
\left\|\boldsymbol{\psi}^J_{k,\ell}\right\|_1 &> \left(\sum_{m=1}^{\ell+1} \xi - (m-1)\overline{\Delta}\right) - \zeta - \left|\mathbf{\Phi}_{k,j^k_{\ell+1}}\right| \\
&= \left(\sum_{m=1}^{\ell} \xi - (m-1)\overline{\Delta}\right) + \xi - \ell\overline{\Delta} - \zeta - \left|\mathbf{\Phi}_{k,j^k_{\ell+1}}\right| \\
&\geq \sum_{m=1}^{\ell} \xi - (m-1)\overline{\Delta},
\end{aligned}$$

where the first inequality follows from $\boldsymbol{\psi}^J_{k,\ell+1} = \boldsymbol{\psi}^J_{k,\ell} + \mathbf{\Phi}_{k,j^k_{\ell+1}}$ and (6.7). The resulting strict inequality contradicts (6.6) and the assertion follows. □

The last preparatory lemmata for the proof of (6.5) establishes two upper bounds on the sum of the largest entries in $\mathbf{\Phi}_{k,\cdot}$ with respect to $\mathbf{\Phi}_{k+1,\cdot}$.

Lemma 6.23 ([82, Lem. 4.7]). *Let Assumption 6.19 hold. Let $k \in \mathbb{N}$, let i^* be the rounding index in the k-th cell and let $i^* \in J \subset \{1,\dots,M\}$.*

Let $[\mathbf{\Phi}]^+_{k,i} > 0$ for all $i \in J$. Then,

$$\left\|\boldsymbol{\psi}^+_{k-1,|J|}\right\|_1 \geq \sum_{i \in J} [\mathbf{\Phi}]^+_{k,i}\,.$$

Let $[\mathbf{\Phi}]^-_{k,i} > 0$ for all $i \in J$. Then,

$$\left\|\boldsymbol{\psi}^-_{k-1,|J|}\right\|_1 \geq \sum_{i \in J} [\mathbf{\Phi}]^-_{k,i} - \overline{\Delta}.$$

Proof. The assumption $[\mathbf{\Phi}]^{\pm}_{k,i} > 0$ for all $i \in J$ ensures that all steps are well-defined, i.e. all summands in the formulas exist.

For the first estimate, we observe that the largest $|J|$ entries in $[\mathbf{\Phi}]^{+}_{k-1,\cdot}$ have to sum to a higher value than the entries in J. This and the fact that i^* was selected as rounding index give

$$\begin{aligned}\left\|\boldsymbol{\psi}^{+}_{k-1,|J|}\right\|_1 &\geq \sum_{i\in J\setminus\{i^*\}} \mathbf{\Phi}_{k-1,i} + \mathbf{\Phi}_{k-1,i^*} \\ &= \sum_{i\in J\setminus\{i^*\}} \left([\mathbf{\Phi}]^{+}_{k,i} - \boldsymbol{\alpha}_{k,i}h_k\right) + [\mathbf{\Phi}]^{+}_{k,i^*} - \boldsymbol{\alpha}_{k,i^*}h_k + h_k \\ &\geq \sum_{i\in J} [\mathbf{\Phi}]^{+}_{k,i}\,.\end{aligned}$$

The last inequality follows from Lemma 6.5. The derivation of the second estimate is similar, but the sign in front of $\boldsymbol{\alpha}_{k,\cdot}$ and $\boldsymbol{\omega}_{k,\cdot}$ switches in the subtraction, which gives the additional summand $-\overline{\Delta}$ in the estimate.

$$\begin{aligned}\left\|\boldsymbol{\psi}^{-}_{k-1,|J|}\right\|_1 &\geq \sum_{i\in J\setminus\{i^*\}} [\mathbf{\Phi}]^{-}_{k-1,i} - \mathbf{\Phi}_{k-1,i^*} \\ &= \sum_{i\in J\setminus\{i^*\}} \left([\mathbf{\Phi}]^{-}_{k,i} + \boldsymbol{\alpha}_{k,i}h_k\right) - (\mathbf{\Phi}_{k,i^*} - \boldsymbol{\alpha}_{k,i^*}h_k + h_k) \\ &= \sum_{i\in J\setminus\{i^*\}} \left([\mathbf{\Phi}]^{-}_{k,i} + \boldsymbol{\alpha}_{k,i}h_k\right) + [\mathbf{\Phi}]^{-}_{k,i^*} + \boldsymbol{\alpha}_{k,i^*}h_k - h_k \\ &\geq \sum_{i\in J} [\mathbf{\Phi}]^{-}_{k,i} - h_k \\ &\geq \sum_{i\in J} [\mathbf{\Phi}]^{-}_{k,i} - \overline{\Delta}.\end{aligned}$$

We make two comments on the first inequality. First, let $\mathbf{\Phi}_{k-1,i} > 0$ and $\mathbf{\Phi}_{k,i} < 0$ for some $i \in J$. Then, $i = i^*$ follows as only one entry can decrease in the k-th cell and consequently $\mathbf{\Phi}_{k-1,i} \leq 0$ for all $i \in J\setminus\{i^*\}$. Second, $\mathbf{\Phi}_{k-1,i^*} \leq 0$ implies $[\mathbf{\Phi}]^{-}_{k-1,i^*} = -\mathbf{\Phi}_{k-1,i^*}$ and in the other case, the inequality holds true as a positive value is subtracted. □

We are ready to prove (6.5) for (SUR-GEN) satisfying Assumption 6.19 which in turn establishes Theorem 6.20. The proof makes use of an inductive argument.

Theorem 6.24 (The first linear bound, Theorem 6.20, [82, Thm 4.8])**.** *Let Assumption 6.19 hold for* (SUR-GEN)*, $\mathfrak{s} \in \{+,-\}$ and $\boldsymbol{\alpha} \in \mathcal{A}$ be fixed. Then,*

$$\sup_{k\in\mathbb{N}} \left\|[\mathbf{\Phi}]^{\mathfrak{s}}_{k,\cdot}\right\|_\infty \leq (M+1)\overline{\Delta}.$$

Proof. For a fixed $C > 0$, we define $k^{\mathfrak{s}}_m$ as the index of the first grid cell, in which the largest m entries of $[\mathbf{\Phi}]^{\mathfrak{s}}_{k,\cdot}$ sum up to a value greater than $\sum_{i=1}^{m}(C-(i-1))\overline{\Delta}$, i.e.

$$k^{\mathfrak{s}}_m(C) \coloneqq \min\left\{k\in\mathbb{N} : \left\|\boldsymbol{\psi}^{\mathfrak{s}}_{k,m}\right\|_1 > \sum_{i=1}^{m}(C-(i-1))\overline{\Delta}\right\}$$

and $k_m^5(C) := \infty$ if the set is empty, i.e. if there is no such index.

First bound: $\sup_k \| [\mathbf{\Phi}]^+_{k,\cdot} \|_\infty \leq M\overline{\Delta}$**.** To this end, we abbreviate $i_\ell^k := i_\ell^{k,+}$ and $k_m := k_m^+(M)$. We proceed with a contradiction argument and assume $\sup_k \| [\mathbf{\Phi}]^+_{k,\cdot} \|_\infty > M\overline{\Delta}$. Thus, there exists $k_1 \in \mathbb{N}$ as defined above such that

$$[\mathbf{\Phi}]^+_{k_1, i_1^{k_1}} > M\overline{\Delta}.$$

Assume we knew that for all $1 \leq m \leq M$ the assertion

$$k_m < \infty \text{ and } k_m < k_{m-1} \text{ if additionally } 2 \leq m \qquad (*)$$

holds. Then, $(*)$ and the definition of k_M yield $k_M < \infty$ and

$$\left\| \boldsymbol{\psi}^+_{k_M, M} \right\|_1 > \sum_{i=1}^{M} (M - (i-1))\overline{\Delta}.$$

Furthermore, $(*)$ yields $k_M < k_{M-1}$, i.e. the sum of the largest M entries of $[\mathbf{\Phi}]^+_{k_M,\cdot}$ exceeds the corresponding bound at an grid cell with a smaller index than the sum of the largest $M-1$ entries exceeds its corresponding bound. Thus,

$$\left\| \boldsymbol{\psi}^+_{k_M, M-1} \right\|_1 \leq \sum_{i=1}^{M-1} (M - (i-1))\overline{\Delta}.$$

Inserting these estimates into Lemma 6.22 implies $[\mathbf{\Phi}]^+_{k_M,m} > 0$ for all $m \in \{1, \dots, M\}$. This contradicts the fact that at least one entry of $\mathbf{\Phi}_{k_M,\cdot}$ has to be strictly negative if there exists a strictly positive entry of $\mathbf{\Phi}_{k_M,\cdot}$, see Lemma 6.5. Thus, the contradictory assumption was wrong and the bound holds for $[\mathbf{\Phi}]^+_{k,\cdot}$ for all $k \in \mathbb{N}$.

We prove $(*)$ inductively. The index k_1 is the smallest k such that $[\mathbf{\Phi}]^+_{k,m} > M\overline{\Delta}$ for some m. Thus, $[\mathbf{\Phi}]^+_{k_1, i_1^{k_1}} > [\mathbf{\Phi}]^+_{k_1-1, m}$ for all $m \in \{1, \dots, M\}$ and in particular,

$$[\mathbf{\Phi}]^+_{k_1, i_1^{k_1}} > [\mathbf{\Phi}]^+_{k_1-1, i_1^{k_1}}.$$

Furthermore, the update formula gives

$$\mathbf{\Phi}_{k_1, i_1^{k_1}} = \mathbf{\Phi}_{k_1-1, i_1^{k_1}} + \boldsymbol{\alpha}_{k_1, i_1^{k_1}} h_{k_1} - \boldsymbol{\omega}_{k_1, i_1^{k_1}} h_{k_1}.$$

We deduce $\boldsymbol{\alpha}_{k_1, i_1^{k_1}} > 0$ as well as $\boldsymbol{\omega}_{k_1, i_1^{k_1}} = 0$ and thus, $i_1^{k_1}$ cannot be the rounding index in grid cell k_1. Let i^* denote the rounding index in grid cell k_1. As $\boldsymbol{\alpha}_{k_1, i_1^{k_1}} > 0$, we have $i_1^{k_1} \in F_{k_1}$. Lemma 6.21 implies $[\mathbf{\Phi}]^+_{k_1, i^*} > (M-1)\overline{\Delta}$ and in turn

$$[\mathbf{\Phi}]^+_{k_1, i_2^{k_1}} > (M-1)\overline{\Delta} \quad \text{and} \quad \left\| \boldsymbol{\psi}^+_{k_1, 2} \right\|_1 > M\overline{\Delta} + (M-1)\overline{\Delta}.$$

Immediately, we deduce $k_2 \leq k_1$. For the grid cell $k_1 - 1$, Lemma 6.23 yields

$$\left\|\boldsymbol{\psi}^+_{k_1-1,2}\right\|_1 \geq [\boldsymbol{\Phi}]^+_{k_1,i_1^{k_1}} + [\boldsymbol{\Phi}]^+_{k_1,i^*} > M\overline{\Delta} + (M-1)\overline{\Delta},$$

which yields $k_2 < k_1$ and concludes the base case for the proof of $(*)$.

Let $2 \leq m \leq M-1$ and assume inductively that $(*)$ holds for m. Thus,

$$\left\|\boldsymbol{\psi}^+_{k_m,m-1}\right\|_1 \leq \sum_{i=1}^{m-1}(M-(i-1))\overline{\Delta} \text{ and}$$
$$\left\|\boldsymbol{\psi}^+_{k_m,m}\right\|_1 > \sum_{i=1}^{m}(M-(i-1))\overline{\Delta}.$$

follow from $k_m < k_{m-1} < \infty$ as in the base case. Again, Lemma 6.22 asserts $[\boldsymbol{\Phi}]^+_{k_m,\ell} > (M-(m-1))\overline{\Delta}$ for all $\ell \in \{i_1^{k_m},\dots,i_m^{k_m}\}$. Let i^* denote the rounding index in grid cell k_m. We deduce $i^* \notin \{i_1^{k_m},\dots,i_m^{k_m}\}$ as $\|\boldsymbol{\psi}^+_{k_m,m}\|_1 > \|\boldsymbol{\psi}^+_{k_m-1,m}\|_1$ by definition of k_m. Similar to the base case, there exists $\ell \in \{i_1^{k_m},\dots,i_m^{k_m}\}$ such that $\boldsymbol{\alpha}_{k_1,\ell} > 0$ due to the increase. Lemma 6.21 yields $[\boldsymbol{\Phi}]^+_{k_m,i^*} > (M-m)\overline{\Delta}$ and in turn $k_{m+1} \leq k_m < \infty$. To see $k_{m+1} < k_m$ and close the induction, we employ Lemma 6.23 and obtain

$$\left\|\boldsymbol{\psi}^+_{k_m-1,m+1}\right\|_1 \geq \sum_{\ell=1}^{m}[\boldsymbol{\Phi}]^+_{k_m,i_\ell^{k_m}} + [\boldsymbol{\Phi}]^+_{k_m,i^*} > \sum_{\ell=1}^{m+1}(M-(\ell-1))\overline{\Delta}.$$

This proves $(*)$ and in turn $\sup_k \|\,[\boldsymbol{\Phi}]^+_{k,\cdot}\,\|_\infty \leq M\overline{\Delta}$.

Second bound: $\sup_k \|\,[\boldsymbol{\Phi}]^-_{k,\cdot}\,\|_\infty \leq (M+1)\overline{\Delta}$. To this end, we abbreviate $i_\ell^k \coloneqq i_\ell^{k,-}$ and $k_m \coloneqq k_m^-(M+1)$. We proceed with a contradiction argument and assume $\sup_k \|\,[\boldsymbol{\Phi}]^-_{k,\cdot}\,\|_\infty > (M+1)\overline{\Delta}$. Thus, there exists $k_1 \in \mathbb{N}$ as defined above such that

$$[\boldsymbol{\Phi}]^-_{k_1,i_1^{k_1}} > (M+1)\overline{\Delta}.$$

Similarly to above, we assume we knew that

for all $2 \leq m \leq M$ at least one of the following two cases holds: $(**)$
(1) $k_m < \infty,$ and $k_m < k_{m-1}$;
(2) $k_m < \infty,$ and $k_{m+1} < \infty,$ and $k_m \leq k_{m-1},\ k_{m+1} < k_{m-1}$.

For $m = M$, case (1) has to hold as only M entries exist. Thus, $k_M < \infty$ and $k_M < k_{M-1}$ and with the same reasoning as above, Lemma 6.22 implies

$$[\boldsymbol{\Phi}]^-_{k_M,j} > 0$$

for all $j \in \{1,\dots,M\}$, which is contradictory to the claim of Lemma 6.5. Thus, the assumption was wrong and the claimed bound holds for $[\boldsymbol{\Phi}]^-_{k,\cdot}$ for all $k \in \mathbb{N}$.

We prove $(**)$ inductively. By definition of k_1, $[\mathbf{\Phi}]^-_{k_1,i_1^{k_1}} > [\mathbf{\Phi}]^-_{k_1-1,j}$ holds for all $j \in \{1,\dots,M\}$ and $i_1^{k_1}$ has to be the rounding index in cell k_1 because at most one entry of $[\mathbf{\Phi}]^-_{k,\cdot}$ may increase in the k-th cell by definition of (SUR-GEN). Furthermore, $[\mathbf{\Phi}]^-_{k_1,i_1^{k_1}} > [\mathbf{\Phi}]^-_{k_1-1,i_1^{k_1}}$ implies $\boldsymbol{\alpha}_{k_1,i_1^{k_1}} < 1$ and Lemma 6.21 implies that existence of some $j \neq i_1^{k_1}$, $j \in F_{k_1}$ such that $[\mathbf{\Phi}]^-_{k_1,j} \geq M\overline{\Delta}$. We deduce

$$\left\|\boldsymbol{\psi}^-_{k_1,2}\right\|_1 > (M+1)\overline{\Delta} + M\overline{\Delta}$$

and $k_2 \leq k_1$. As the rounding index was $i_1^{k_1}$, we infer $\|\boldsymbol{\psi}^-_{k_1-1,2}\|_1 > (M+1)\overline{\Delta} + M\overline{\Delta}$ from Assumption 6.19 if $F_{k_1} = \{i_1^{k_1}, j\}$ and in turn $k_2 < k_1$. If $\{i_1^{k_1}, j\} \subsetneq F_{k_1}$, there exists $\ell \in F_{k_1}\backslash\{i_1^{k_1}, j\}$ with $[\mathbf{\Phi}]^-_{k_1,\ell} \geq M\overline{\Delta}$ by Lemma 6.21. Lemma 6.23 yields

$$\begin{aligned}\left\|\boldsymbol{\psi}^-_{k_1-1,3}\right\|_1 &\geq [\mathbf{\Phi}]^-_{k_1,i_1^{k_1}} + [\mathbf{\Phi}]^-_{k_1,j} + [\mathbf{\Phi}]^-_{k_1,\ell} - \overline{\Delta} \\ &> (M+1)\overline{\Delta} + 2M\overline{\Delta} - \overline{\Delta} = (M+1)\overline{\Delta} + M\overline{\Delta} + (M-1)\overline{\Delta},\end{aligned}$$

which proves the claim for k_1 and concludes the base case.

Let $m \leq M-1$ and assume inductively that $(**)$ holds for m. If case (1) holds, i.e. $k_m < k_{m-1}$, Lemma 6.22 yields

$$[\mathbf{\Phi}]^-_{k_m,i_m^{k_m}} > (M+1-(m-1))\overline{\Delta}.$$

Then, the rounding index is in $\{i_1^{k_m},\dots,i_m^{k_m}\}$ as the sum of the largest m entries of the negative part increased. The estimate $[\mathbf{\Phi}]^-_{k_m,j} \geq (M+1-m)\overline{\Delta}$ holds by virtue of Lemma 6.21 and Assumption 6.19 for some $j \notin \{i_1^{k_m},\dots,i_m^{k_m}\}$, $j \in F_{k_m}$. We deduce $k_{m+1} \leq k_m < \infty$ by definition of k_{m+1} and estimate

$$\|\boldsymbol{\psi}^-_{k_m,m+1}\|_1 \geq \|\boldsymbol{\psi}^-_{k_m,m}\|_1 + [\mathbf{\Phi}]^-_{k_m,j} > \sum_{n=1}^{m+1}(M+1-(n-1))\overline{\Delta}.$$

If $F_{k_m} \subset \{i_1^{k_m},\dots,i_m^{k_m},j\}$, we obtain

$$\begin{aligned}\|\boldsymbol{\psi}^-_{k_m-1,m+1}\|_1 &\geq \sum_{n=1}^{m}[\mathbf{\Phi}]^-_{k_m-1,i_n^{k_m}} + [\mathbf{\Phi}]^-_{k_m-1,j} \\ &= \sum_{n=1}^{m}[\mathbf{\Phi}]^-_{k_m,i_n^{k_m}} + [\mathbf{\Phi}]^-_{k_m,j} \\ &> \sum_{n=1}^{m+1}(M+1-(n-1))\overline{\Delta}\end{aligned}$$

and deduce $k_{m+1} < k_m$. The first inequality is true because $[\mathbf{\Phi}]^-_{k_m,n} > \overline{\Delta}$ holds for all $n \in F_{k_m}$. Thus, $[\mathbf{\Phi}]^-_{k_m-1,n} > 0$ for all $n \in F_{k_m}$ as an entry can change by at most $\overline{\Delta}$ from

one grid cell to the next. If on the other hand there exists $\ell \in F_{k_m} \backslash \{i_1^{k_m}, \dots, i_k^{k_m}, j\}$, an analogous reasoning to the base case above yields $k_{m+2} < k_m$:

$$\begin{aligned}\left\|\boldsymbol{\psi}^-_{k_m-1,m+2}\right\|_1 &\geq \sum_{n=1}^{m} [\boldsymbol{\Phi}]^-_{k_m,i_n^{k_m}} + [\boldsymbol{\Phi}]^-_{k_m,j} + [\boldsymbol{\Phi}]^-_{k_m,\ell} - \overline{\Delta} \\ &> \sum_{n=1}^{m} (M+1-(n-1))\overline{\Delta} + 2(M+1-m)\overline{\Delta} - \overline{\Delta} \\ &= \sum_{n=1}^{m+2} (M+1-(n-1))\overline{\Delta},\end{aligned}$$

where we have employed Lemma 6.23 again. Note that this is well-defined for $m = M-1$ because this implies $F_{k_m} \subset \{i_1^{k_m}, \dots, i_m^{k_m}, j\}$.

Let case (2) hold. If $[\boldsymbol{\Phi}]^-_{k_{m+1},i_{m+1}^{k_{m+1}}} \leq (M+1-m)\overline{\Delta}$, we deduce

$$\begin{aligned}\left\|\boldsymbol{\psi}^-_{k_{m+1},m}\right\|_1 &> \sum_{i=1}^{m+1} (M+1-(i-1))\overline{\Delta} - [\boldsymbol{\Phi}]^-_{k_{m+1},i_{m+1}^{k_{m+1}}} \\ &\geq \sum_{i=1}^{m} (M+1-(i-1))\overline{\Delta},\end{aligned}$$

which implies $k_m \leq k_{m+1}$. Inserting the induction hypothesis of case (2), we get $k_m \leq k_{m+1} < k_{m-1}$. Hence, case (1) holds and the reasoning has already been handled. Thus, we restrict to the case, in which the estimates

$$[\boldsymbol{\Phi}]^-_{k_{m+1},i_m^{k_{m+1}}} > (M+1-m)\overline{\Delta} \quad \text{and} \quad [\boldsymbol{\Phi}]^-_{k_{m+1},i_{m+1}^{k_{m+1}}} > (M+1-m)\overline{\Delta}$$

hold. A reasoning similar to the argument in Lemma 6.22 and the induction hypothesis $k_{m+1} < k_{m-1}$ of case (2) also gives the estimate

$$[\boldsymbol{\Phi}]^-_{k_{m+1},i_m^{k_{m+1}}} + [\boldsymbol{\Phi}]^-_{k_{m+1},i_{m+1}^{k_{m+1}}} > (M+1-(m-1))\overline{\Delta} + (M+1-m)\overline{\Delta}.$$

As the sum of the values of the largest $m+1$ negative entries increased by definition of k_{m+1}, the rounding index has to be in $\{i_1^{k_{m+1}}, \dots, i_{m+1}^{k_{m+1}}\}$. From Lemma 6.21 and Assumption 6.19, we infer $[\boldsymbol{\Phi}]^-_{k_{m+1},j} \geq (M+1-(m+1))\overline{\Delta}$ for some $j \notin \{i_1^{k_{m+1}}, \dots, i_{m+1}^{k_{m+1}}\}$, $j \in F_{k_{m+1}}$. This cannot happen if $m = M-1$ and then case (1) applies again. As above, we deduce $k_{m+2} \leq k_{m+1} < \infty$. If $F_{k_{m+1}} \subset \{i_1^{k_{m+1}}, \dots, i_{m+1}^{k_{m+1}}, j\}$, we deduce $k_{m+2} < k_{m+1}$ as above. Otherwise, there exists $\ell \in F_{k_{m+1}} \backslash \{i_1^{k_{m+1}}, \dots, i_{m+1}^{k_{m+1}}, j\}$ and Lemmas 6.21 and 6.23 yield

$$\begin{aligned}\left\|\boldsymbol{\psi}^-_{k_{m+1}-1,m+3}\right\|_1 &\geq \sum_{n=1}^{m+1} [\boldsymbol{\Phi}]^-_{k_{m+1},i_n^{k_{m+1}}} + [\boldsymbol{\Phi}]^-_{k_{m+1},j} + [\boldsymbol{\Phi}]^-_{k_{m+1},\ell} - \overline{\Delta} \\ &> \sum_{m=1}^{m+1} (M+1-(n-1))\overline{\Delta} + 2(M+1-(m+1))\overline{\Delta} - \overline{\Delta} \\ &= \sum_{n=1}^{m+3} (M+1-(n-1))\overline{\Delta},\end{aligned}$$

which gives $k_{m+3} < k_{m+1}$. Note that this is well-defined for $m = M-2$ as this implies $F_{k_{m+1}} \subset \{i_1^{k_{m+1}}, \dots, i_{m+1}^{k_{m+1}}, j\}$, which yields $k_{m+2} < k_{m+1}$ and $m = M-1$ has been excluded before. This closes the induction proving $(**)$. □

6.5 An asymptotically tight bound for (SUR-VC)

The following results have been established in Sections 3, 4.3, 4.4 and 5 in the article [82] by Kirches, Lenders, and the author. We estimate the integrality gap as follows.

Theorem 6.25 ([82, Prop. 3.6])**.** *Let $\boldsymbol{\omega}$ be a binary control computed by* (SUR-VC) *from a relaxed control $\boldsymbol{\alpha}$ on a rounding grid with mesh size $\overline{\Delta}$. Then,*

$$d(\boldsymbol{\omega}, \boldsymbol{\alpha}) \leq \left\lfloor \frac{M}{2} \right\rfloor \overline{\Delta}.$$

Remark 6.26. *This bound is asymptotically tight. To see this, we refer to an example in the supplementary material of* [68]*, which constructs a relaxed control $\boldsymbol{\alpha}$ iteratively on an equidistant grid such that $d(\boldsymbol{\omega}, \boldsymbol{\alpha}) = \frac{M-1}{2}\overline{\Delta}$ if $\boldsymbol{\omega}$ is computed with* (SUR-VC)*.*

Outline of proof. We begin with technical assumptions. In particular, we assume that there always exists a relaxed control $\boldsymbol{\alpha}$ that transforms the integrated control deviation into a so-called ε-stairs-shaped form. Then, we prove the theorem with an inductive strategy that relies on this ability. Finally, we provide algorithms and proofs to obtain the transformation constructively.

6.5.1 Technical assumptions

The following two preparations are necessary for technical constructions to prove the sharper bound. Again, we extend the sequence $(\mathcal{T}_k)_{k\in\{1,\dots,N\}}$ to $(\mathcal{T}_k)_{k\in\mathbb{N}}$, on which we impose the following assumption.

Assumption 6.27 (Recurrence of $\overline{\Delta}$ in $(h_k)_{k\in\mathbb{N}}$)**.** *The sequence of cell volumes $(h_k)_{k\in\mathbb{N}}$ has a subsequence $(h_{k_\ell})_{\ell\in\mathbb{N}}$ with $h_{k_\ell} \equiv \overline{\Delta}$.*

Remark 6.28. *Assumption 6.27 does not restrict the applicability of Theorem 6.25 as the first element of $(h_{k_\ell})_{\ell\in\mathbb{N}}$ may occur after the* real *grid cells, i.e. $k_1 > N$.*

Definition 6.29 (ε-stairs-shape, [82, Def. 4.11])**.** *Let $k \in \mathbb{N}$, $\mathfrak{s} \in \{-,+\}$ and $\varepsilon > 0$. We call $\boldsymbol{\Phi}_{k,\cdot}$ ε-**stairs-shaped** in a set $J \coloneqq \{j_1, \dots, j_{|J|}\} \subset I_k^{\mathfrak{s}}$ if there exists $1 \leq m \leq |J|$ such that for all $1 \leq i \leq m$*

$$\left| [\boldsymbol{\Phi}_{k,j_i}]^{\mathfrak{s}} - [\boldsymbol{\Phi}]^{\mathfrak{s}}_{k,j_{i+1}} \right| \in B_\varepsilon(\overline{\Delta}) \tag{6.8}$$

and for all $m+1 \leq i \leq |J|$

$$\left| [\boldsymbol{\Phi}]^{\mathfrak{s}}_{k,j_i} \right| \in [0, \overline{\Delta}). \tag{6.9}$$

For a fixed $k_1 \in \mathbb{N}$, the integrated control deviation vector $\mathbf{\Phi}_{k_1,\cdot}$ does not depend on $(\boldsymbol{\alpha}_{k,\cdot})_{k>k_1}$. Our proof makes use of the fact that for all $\mathbf{\Phi}_{k_1,\cdot}$, we can construct a finite subsequence $(\boldsymbol{\alpha}_{k,\cdot})_{k_1 \leq k \leq k_2}$ of $(\boldsymbol{\alpha}_{k,\cdot})_{k\in\mathbb{N}}$ such that the application of (SUR-VC) implies

1. $[\mathbf{\Phi}]^{\mathfrak{s}}_{k_2,\cdot}$ is ε-*stairs-shaped* and
2. $\|\mathbf{\Phi}_{k_1,\cdot}\|_1 = \|\mathbf{\Phi}_{k_1+1,\cdot}\|_1 = \ldots = \|\mathbf{\Phi}_{k_2,\cdot}\|_1$.

The existence of such a sequence is not obvious. We prove the existence constructively in Section 6.5.3 and state it here such that the proof of Theorem 6.25 can work.

Proposition 6.30. *Let $0 < \varepsilon < \frac{\overline{\Delta}}{2}$. Let the assumptions of Theorem 6.25 hold. Let the extension $(\mathcal{T}_k)_{k\in\mathbb{N}}$ of the sequence of grid cells satisfy Assumption 6.27. Let $\mathbf{\Phi}_{k_1,\cdot}(\boldsymbol{\alpha})$ be the control deviation after grid cell k_1. Let $\mathfrak{s} \in \{+,-\}$. Then, there exists $\boldsymbol{\beta} \in \mathcal{A}$ such that*

1. *$\boldsymbol{\beta}_{k,\cdot} = \boldsymbol{\alpha}_{k,\cdot}$ for $k \leq k_1$,*
2. *there exists $k_2 \geq k_1$ such that $\mathbf{\Phi}_{k_2,\cdot}(\boldsymbol{\beta})$ is ε-stairs-shaped in $J \subset I^{\mathfrak{s}}_{k_2} = I^{\mathfrak{s}}_{k_1}$,*
3. *$\mathbf{\Phi}_{k_1,j}(\boldsymbol{\beta}) = \mathbf{\Phi}_{k_1+1,j}(\boldsymbol{\beta}) = \ldots = \mathbf{\Phi}_{k_2,j}(\boldsymbol{\beta})$ for $j \notin J$ and*
4. *$\|\boldsymbol{\psi}^J_{k_1,|J|}(\boldsymbol{\beta})\|_1 = \|\boldsymbol{\psi}^J_{k_1+1,|J|}(\boldsymbol{\beta})\|_1 = \ldots = \|\boldsymbol{\psi}^J_{k_2,|J|}(\boldsymbol{\beta})\|_1$.*

Proof. The claim will follow from Lemma 6.43. □

The first claim asserts that $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ induce an identical behavior of (SUR-VC), in particular identical control deviations, up to the k_1-th grid cell. The claims 2-4 give that we can choose $\boldsymbol{\beta}_{k_1+1}, \ldots, \boldsymbol{\beta}_{k_2}$ such that

1. $\mathbf{\Phi}_{k_2,\cdot}$ is ε-stairs-shaped in a fixed subset J of either positive or negative entries,
2. the entries of $\mathbf{\Phi}_{k,\cdot}$ that are not contained in J are not affected over the cells $k \in \{k_1, \ldots, k_2\}$ and
3. the sum-norm of the affected entries is conserved over the cells $k \in \{k_1, \ldots, k_2\}$.

The last two claims are equivalent as J contains only indices of entries of either the positive part or the negative part. However, we have made this explicit to improve the accessibility of the argument.

6.5.2 Tightening the bound

We abbreviate $i^k_j := i^{k,\mathfrak{s}}_j$ if no ambiguity is present.

Definition 6.31 ([82, Def. 4.13])**.** *We define the quantities to bound for $\mathfrak{s} \in \{+,-\}$ as*

$$S^{\mathfrak{s}}_1 := \sup_{\boldsymbol{\alpha}\in\mathcal{A}} \sup_{k\in\mathbb{N}} \left\| [\mathbf{\Phi}]^{\mathfrak{s}}_{k,\cdot} \right\|_\infty = \sup_{\boldsymbol{\alpha}\in\mathcal{A}} \sup_{k\in\mathbb{N}} \left\| \boldsymbol{\psi}^{\mathfrak{s}}_{k,1} \right\|_1 .$$

S_1^+ and S_1^- are bounded and well-defined by Theorem 6.24. To state claims on parts of $\mathbf{\Phi}_{k,\cdot}$, we introduce the following sets of cell indices.

Definition 6.32 ([82, Def. 4.14]). *For $\ell \in \{1,\dots,M\}$ and $\mathfrak{s} \in \{+,-\}$, we define the set of grid cells k for which at least ℓ strictly negative (positive) entries of $\mathbf{\Phi}_{k,\cdot}$ exist*

$$N_\ell^{\mathfrak{s}} \coloneqq \{k \in \mathbb{N} : \ell \leq |I_k^{\mathfrak{s}}| \text{ and } [\mathbf{\Phi}]^{\mathfrak{s}}_{k,i_\ell^{k,\mathfrak{s}}} > 0\}.$$

Obviously, the inclusion $N_{\ell+1}^{\mathfrak{s}} \subset N_\ell^{\mathfrak{s}}$ holds. The following claims hold for fixed $\boldsymbol{\alpha} \in \mathcal{A}$ and $\mathbf{\Phi}(\boldsymbol{\alpha})$, $\boldsymbol{\psi}^{\mathfrak{s}}(\boldsymbol{\alpha})$ resulting from the application of (SUR-VC). We omit $\boldsymbol{\alpha}$ in most of the reasoning for sake of brevity. Assuming $N_\ell \neq \emptyset$ for ε, we can bound the sum of the ℓ largest entries of the integrated control deviation vector $[\mathbf{\Phi}]^{\mathfrak{s}}_{k,\cdot}$ from above for $k \in N_\ell$.

Lemma 6.33 ([82, Lem. 4.15]). *Let Assumption 6.27 hold, $\mathfrak{s} \in \{+,-\}$, $L \coloneqq \left\lfloor \frac{S_1^{\mathfrak{s}}}{\overline{\Delta}} \right\rfloor$ and $N_L^{\mathfrak{s}} \neq \emptyset$. Then, for all $\ell \in \{1,\dots,L\}$, we have*

$$\sup_{k \in N_\ell} \|\boldsymbol{\psi}^{\mathfrak{s}}_{k,\ell}\|_1 \leq \sum_{i=1}^{\ell} S_1^{\mathfrak{s}} - (i-1)\overline{\Delta}.$$

Proof. By definition of $S_1^{\mathfrak{s}}$, the claim holds for $\ell = 1$. We proceed inductively and assume the claim holds for $\ell \in \mathbb{N}$ with $\ell + 1 \leq L$. We argue by contradiction and suppose the claim does not hold true for $\ell+1$, i.e.

$$\sup_{k \in N_{\ell+1}} \|\boldsymbol{\psi}^{\mathfrak{s}}_{k,\ell+1}\|_1 > \sum_{i=1}^{\ell+1} S_1^{\mathfrak{s}} - (i-1)\overline{\Delta}. \tag{6.10}$$

We set

$$d \coloneqq \sup_{k \in N_{\ell+1}} \|\boldsymbol{\psi}^{\mathfrak{s}}_{k,\ell+1}\|_1 - \left(\sum_{i=1}^{\ell+1} S_1^{\mathfrak{s}} - (i-1)\overline{\Delta} \right).$$

Thus, for all $0 < \varepsilon < d$, there exist $k_1 \in \mathbb{N}$ such that

$$\|\boldsymbol{\psi}^{\mathfrak{s}}_{k_1,\ell+1}\|_1 = \left(\sum_{i=1}^{\ell+1} S_1^{\mathfrak{s}} - (i-1)\overline{\Delta} \right) + (d - \varepsilon). \tag{6.11}$$

We recall that $\mathbf{\Phi}_{k_1,\cdot}$ depends only on $(\boldsymbol{\alpha}_{k,\cdot})_{k \leq k_1}$ and the same holds for the $\boldsymbol{\beta} \in \mathcal{A}$ provided by Proposition 6.30. For $\delta > 0$ small enough, Proposition 6.30 also gives $k_2 \coloneqq k_2(\delta) \geq k_1$ such that

$$\|\boldsymbol{\psi}^J_{k_1,\ell+1}(\boldsymbol{\alpha})\|_1 = \|\boldsymbol{\psi}^J_{k_1,\ell+1}(\boldsymbol{\beta})\|_1 = \|\boldsymbol{\psi}^J_{k_1+1,\ell+1}(\boldsymbol{\beta})\|_1 = \ldots = \|\boldsymbol{\psi}^J_{k_2,\ell+1}(\boldsymbol{\beta})\|_1$$

and $\mathbf{\Phi}_{k_2,\cdot}(\boldsymbol{\beta})$ is δ-stairs-shaped in $J \coloneqq \{i_1^{k_1},\dots,i_{\ell+1}^{k_1}\}$ comprising the largest $\ell+1$ entries of $[\mathbf{\Phi}]^{\mathfrak{s}}_{k_1,\cdot}$. We apply Lemma 6.22 to the estimate (6.10) and the induction hypothesis,

which yields the estimate $[\mathbf{\Phi}]^{\mathfrak{s}}_{k,j}(\boldsymbol{\beta}) > \overline{\Delta}$ for all $j \in J$ and all $k_1 \leq k \leq k_2$. Thus, the case (6.8) is always present in Proposition 6.30. Driving $\delta \to 0$, we obtain

$$[\mathbf{\Phi}]^{\mathfrak{s}}_{k_2,j_1^{k_2}}(\boldsymbol{\beta}) \to S_1^{\mathfrak{s}} + \frac{d-\varepsilon}{\ell+1},$$

$$\dots$$

$$[\mathbf{\Phi}]^{\mathfrak{s}}_{k_2,j_{\ell+1}^{k_2}}(\boldsymbol{\beta}) \to S_1^{\mathfrak{s}} - \ell\overline{\Delta} + \frac{d-\varepsilon}{\ell+1}.$$

As $d-\varepsilon > 0$, we obtain $[\mathbf{\Phi}]^{\mathfrak{s}}_{k_2,j_1^{k_2}} > S_1^{\mathfrak{s}}$ for some $\delta > 0$. This contradicts the definition of $S_1^{\mathfrak{s}}$ and consequently, the estimate (6.10) cannot hold, which closes the argument. □

We bound the same supremum from below with an inductive argument. We provide the induction step as an explicit lemma.

Lemma 6.34 ([82, Lem. 4.16]). *Let Assumption 6.27 hold, $0 < \varepsilon < \overline{\Delta}$, $\mathfrak{s} \in \{+,-\}$, $L \coloneqq \left\lfloor \frac{S_1^{\mathfrak{s}}}{\overline{\Delta}} \right\rfloor$, $\ell+1 \leq L$, $N_\ell^{\mathfrak{s}} \neq \emptyset$ and define*

$$k_\ell^\varepsilon \coloneqq \min\left\{ k \in N_\ell^{\mathfrak{s}} : \sup_{\kappa \in N_\ell^{\mathfrak{s}}} \|\boldsymbol{\psi}^{\mathfrak{s}}_{\kappa,\ell}\|_1 - \|\boldsymbol{\psi}^{\mathfrak{s}}_{k,\ell}\|_1 < \varepsilon \right\}.$$

Furthermore, let the estimate

$$\|\boldsymbol{\psi}^{\mathfrak{s}}_{k_\ell^\varepsilon,\ell}\|_1 > \sum_{i=1}^{\ell} S_1^{\mathfrak{s}} - (i-1)\overline{\Delta} - \varepsilon \tag{6.12}$$

hold. Then, $k_\ell^\varepsilon \in N_{\ell+1}^{\mathfrak{s}} \neq \emptyset$ and the following inequalities hold

$$\|\boldsymbol{\psi}^{\mathfrak{s}}_{k_{\ell+1}^{2\varepsilon},\ell+1}\|_1 > \sum_{i=1}^{\ell+1} S_1^{\mathfrak{s}} - (i-1)\overline{\Delta} - 2\varepsilon, \tag{6.13}$$

$$\|\boldsymbol{\psi}^{\mathfrak{s}}_{k_\ell^\varepsilon,\ell+1}\|_1 > \sum_{i=1}^{\ell+1} S_1^{\mathfrak{s}} - (i-1)\overline{\Delta} - 2\varepsilon,$$

$$[\mathbf{\Phi}]^{\mathfrak{s}}_{k_\ell^\varepsilon,i_\ell^{k_\ell^\varepsilon}} \geq S_1^{\mathfrak{s}} - (\ell-1)\overline{\Delta} - \varepsilon,$$

$$k_{\ell+1}^{2\varepsilon} \leq k_\ell^\varepsilon.$$

Proof. We apply Lemma 6.33, which yields the estimate

$$\sup_{\kappa \in N_{\ell-1}} \|\boldsymbol{\psi}^{\mathfrak{s}}_{\kappa,\ell-1}\|_1 \leq \sum_{i=1}^{\ell-1} S_1^{\mathfrak{s}} - (i-1)\overline{\Delta}.$$

Furthermore, $k_\ell^\varepsilon \in N_\ell^{\mathfrak{s}}$ implies $k_\ell^\varepsilon \in N_1^{\mathfrak{s}}, \dots, N_{\ell-1}^{\mathfrak{s}}$. Combining this estimate with (6.12) and applying Lemma 6.22 gives

$$[\mathbf{\Phi}]^{\mathfrak{s}}_{k_\ell^\varepsilon,i_\ell^{k_\ell^\varepsilon}} > S_1^{\mathfrak{s}} - (\ell-1)\overline{\Delta} - \varepsilon \underset{\ell+1\leq L}{\geq} 2\overline{\Delta} - \varepsilon > \overline{\Delta}. \tag{6.14}$$

We deduce $[\mathbf{\Phi}]^{\mathfrak{s}}_{k_\ell^\varepsilon-1,j} > 0$ for all $j \in \{i_1^{k_\ell^\varepsilon}, \dots, i_\ell^{k_\ell^\varepsilon}\}$. By definition of k_ℓ^ε as the first index such that the estimate holds, we observe that at least one entry of $[\mathbf{\Phi}]^{\mathfrak{s}}_{k_\ell^\varepsilon,\cdot}$ was increased compared to $[\mathbf{\Phi}]^{\mathfrak{s}}_{k_\ell^\varepsilon-1,\cdot}$. Now, we distinguish the cases $\mathfrak{s} = +$ and $\mathfrak{s} = -$.

Case $\mathfrak{s} = -$. We deduce that the rounding index i^* is in $\left\{i_1^{k_\ell^\varepsilon}, \dots, i_\ell^{k_\ell^\varepsilon}\right\}$. As the sum of these entries, which were also all strictly negative before, changed, there exists $j \in F_{k_\ell^\varepsilon} \setminus \left\{i_1^{k_\ell^\varepsilon}, \dots, i_\ell^{k_\ell^\varepsilon}\right\}$. The application of Lemma 6.21 yields

$$\boldsymbol{\Phi}_{k_\ell^\varepsilon, j} - \boldsymbol{\Phi}_{k_\ell^\varepsilon, i^*} \leq \overline{\Delta}. \tag{6.15}$$

with $\boldsymbol{\Phi}_{k_\ell^\varepsilon, j} < 0$ by (6.14), which also yields

$$[\boldsymbol{\Phi}]^-_{k_\ell^\varepsilon, j} \underset{(6.15)}{\geq} [\boldsymbol{\Phi}]^-_{k_\ell^\varepsilon, i^*} - \overline{\Delta} \underset{(6.14)}{>} S_1^- - \ell\overline{\Delta} - \varepsilon \underset{\ell+1 \leq L}{\geq} \overline{\Delta} - \varepsilon > 0.$$

Case $\mathfrak{s} = +$. We deduce that the rounding index i^* is not in $\left\{i_1^{k_\ell^\varepsilon}, \dots, i_\ell^{k_\ell^\varepsilon}\right\}$ and there exists some $j \in F_{k_\ell^\varepsilon} \cap \left\{i_1^{k_\ell^\varepsilon}, \dots, i_\ell^{k_\ell^\varepsilon}\right\}$ as the sum of these strictly positive entries increased. The application of Lemma 6.21 yields

$$\boldsymbol{\Phi}_{k_\ell^\varepsilon, j} - \boldsymbol{\Phi}_{k_\ell^\varepsilon, i^*} \leq \overline{\Delta} \tag{6.16}$$

with $\boldsymbol{\Phi}_{k_\ell^\varepsilon, j} > 0$ by (6.14), which also yields

$$[\boldsymbol{\Phi}]^+_{k_\ell^\varepsilon, i^*} \underset{(6.16)}{\geq} [\boldsymbol{\Phi}]^+_{k_\ell^\varepsilon, j} - \overline{\Delta} \underset{(6.14)}{>} S_1^+ - \ell\overline{\Delta} - \varepsilon \underset{\ell+1 \leq L}{\geq} \overline{\Delta} - \varepsilon > 0.$$

In both cases, we obtain $k_\ell^\varepsilon \in N_{\ell+1}^{\mathfrak{s}}$. We choose $\iota = j$ for the case $\mathfrak{s} = -$ and $\iota = i^*$ for the case $\mathfrak{s} = +$ and obtain the estimates

$$\begin{aligned}
\|\boldsymbol{\psi}^{\mathfrak{s}}_{k_\ell^\varepsilon, \ell+1}\|_1 &\geq \|\boldsymbol{\psi}^{\mathfrak{s}}_{k_\ell^\varepsilon, \ell}\|_1 + [\boldsymbol{\Phi}]^{\mathfrak{s}}_{k_\ell^\varepsilon, \iota} \\
&> \sum_{i=1}^{\ell} S_1^{\mathfrak{s}} - (i-1)\overline{\Delta} - \varepsilon + S_1^{\mathfrak{s}} - \ell\overline{\Delta} - \varepsilon \\
&\geq \sum_{i=1}^{\ell+1} S_1^{\mathfrak{s}} - (i-1)\overline{\Delta} - 2\varepsilon.
\end{aligned}$$

Consequently, we get $k_{\ell+1}^{2\varepsilon} \leq k_\ell^\varepsilon$ and

$$\|\boldsymbol{\psi}^{\mathfrak{s}}_{k_{\ell+1}^{2\varepsilon}, \ell+1}\|_1 > \sum_{i=1}^{\ell+1} S_1^{\mathfrak{s}} - (i-1)\overline{\Delta} - 2\varepsilon.$$

□

The lower bound can be obtained inductively using Lemma 6.34 for the step.

Lemma 6.35 ([82, Lem. 4.17]). *Let Assumption 6.27 hold, $0 < \varepsilon < \overline{\Delta}$, $L := \left\lfloor \frac{S_1^{\mathfrak{s}}}{\overline{\Delta}} \right\rfloor$. There exists $\boldsymbol{\alpha} \in \mathcal{A}$ such that for all $\ell \in \{1, \dots, L\}$ and for $\varepsilon_\ell := \frac{\varepsilon}{2^{L-\ell}}$, we have*

$$\sup_{k \in N_\ell} \|\boldsymbol{\psi}^{\mathfrak{s}}_{k,\ell}\|_1 > \sum_{i=1}^{\ell} S_1^{\mathfrak{s}} - (i-1)\overline{\Delta} - \varepsilon_\ell.$$

Proof. By definition of the sup, there exists $\boldsymbol{\alpha} \in \mathcal{A}$ such that we have $N_1^{\mathfrak{s}} \neq \emptyset$ and $\sup_{k \in N_1^{\mathfrak{s}}} \|\boldsymbol{\psi}_{k,1}^{\mathfrak{s}}\|_1 > S_1^{\mathfrak{s}} - \varepsilon_1$. We proceed inductively to verify the claim for $\ell + 1$ if $\ell + 1 \leq L$ and assume it holds for some ℓ. To this end, we employ Lemma 6.34, in particular (6.13), with the choice $\varepsilon = \varepsilon_\ell$. This gives $N_{\ell+1}^{\mathfrak{s}} \neq \emptyset$ and the existence of $k_{\ell+1}^{\varepsilon_{\ell+1}} \in \mathbb{N}$ such that the estimate

$$\sup_{k \in N_{\ell+1}^{\mathfrak{s}}} \|\boldsymbol{\psi}_{k,\ell+1}^{\mathfrak{s}}\|_1 \geq \|\boldsymbol{\psi}_{k_{\ell+1}^{\varepsilon_{\ell+1}},\ell+1}^{\mathfrak{s}}\|_1 \underset{(6.13)}{>} \sum_{i=1}^{\ell+1} S_1^{\mathfrak{s}} - (i-1)\overline{\Delta} - \varepsilon_{\ell+1}$$

is satisfied, which proves the claim. □

We summarize our insights from Lemmas 6.33 to 6.35 in the following theorem.

Theorem 6.36 ([82, Thm 4.19])**.** *Let Assumption 6.27 hold, let* $\mathfrak{s} \in \{-,+\}$, $L^{\mathfrak{s}} \coloneqq \left\lfloor \frac{S_1^{\mathfrak{s}}}{\overline{\Delta}} \right\rfloor$ *and* $\ell \in \{1, \dots, L^{\mathfrak{s}}\}$*. Then,*

$$\sup_{\boldsymbol{\alpha} \in \mathcal{A}} \sup_{k \in N_\ell^{\mathfrak{s}}} \|\boldsymbol{\psi}_{k,\ell}^{\mathfrak{s}}\|_1 = \sum_{i=1}^{\ell} S_1^{\mathfrak{s}} - (i-1)\overline{\Delta}$$

with $\sup_{k \in \emptyset} \|\boldsymbol{\psi}_{k,\ell}^{\mathfrak{s}}\|_1 \coloneqq 0$.

Theorem 6.36 implies the following corollary, which establishes Theorem 6.25.

Corollary 6.37. *Let Assumption 6.27 hold. Let* $\mathfrak{s} \in \{-,+\}$*. Then,*

$$S_1^{\mathfrak{s}} \leq \left\lfloor \frac{M}{2} \right\rfloor \overline{\Delta}.$$

Proof. We define $L^{\mathfrak{s}} \coloneqq \left\lfloor S_1^{\mathfrak{s}} / \overline{\Delta} \right\rfloor$ for $\mathfrak{s} \in \{-,+\}$. Let $\mathfrak{s} \in \{-,+\}$ and assume that the converse estimate

$$S_1^{\mathfrak{s}} > \left(\left\lfloor \frac{M}{2} \right\rfloor + c \right) \overline{\Delta} \tag{6.17}$$

holds for some $c > 0$, i.e. $L^{\mathfrak{s}} \geq \lfloor M/2 \rfloor$. Theorem 6.36 implies that for all $\zeta > 0$, in particular $\zeta < c\overline{\Delta}$, there exists $\boldsymbol{\alpha} \in \mathcal{A}$ with a minimal $k_\zeta \in \mathbb{N}$ such that

$$\|\boldsymbol{\psi}_{k_\zeta,\lfloor M/2 \rfloor}^{\mathfrak{s}}\|_1 > \sum_{i=1}^{\lfloor M/2 \rfloor} S_1^{\mathfrak{s}} - (i-1)\overline{\Delta} - \zeta$$

and

$$\|\boldsymbol{\psi}_{k_\zeta,\lfloor M/2 \rfloor - 1}^{\mathfrak{s}}\|_1 \leq \sum_{i=1}^{\lfloor M/2 \rfloor - 1} S_1^{\mathfrak{s}} - (i-1)\overline{\Delta}.$$

We employ Lemma 6.22 and deduce

$$[\boldsymbol{\Phi}]_{k_\zeta, i_{\lfloor M/2 \rfloor}}^{k_{\zeta,\mathfrak{s}}, \mathfrak{s}} > \overline{\Delta} + c\overline{\Delta} - \zeta \underset{\zeta < c\overline{\Delta}}{>} \overline{\Delta}.$$

As $k_\zeta \in \mathbb{N}$ is minimal, we use the reasoning from the proof of Lemma 6.34 to deduce the existence of $j \notin \left\{i_1^{k_\zeta,\mathfrak{s}}, \dots, i_{\lfloor M/2 \rfloor}^{n_\zeta,\mathfrak{s}}\right\}$ that satisfies

$$[\mathbf{\Phi}]^{\mathfrak{s}}_{k_\zeta,j} > 0,$$

in particular $N^{\mathfrak{s}}_{\lfloor M/2 \rfloor+1} \neq \emptyset$. For $\mathfrak{s} = +$, let $\mathfrak{r} = -$ and for $\mathfrak{s} = -$, let $\mathfrak{r} = +$. With this notation, we have $|I_k^{\mathfrak{rr}}| \leq \lfloor M/2 \rfloor$ for all $k \in N^{\mathfrak{s}}_{\lfloor M/2 \rfloor+1}$, in particular for $k = k_\zeta$. Let $L^{\mathfrak{r}} \geq |I_{k_\zeta}^{\mathfrak{rr}}|$. Then, Lemma 6.5 and Theorem 6.36 give the estimate

$$\sum_{i=1}^{\lfloor M/2 \rfloor} S_1^{\mathfrak{s}} - (i-1)\overline{\Delta} < \left\| \boldsymbol{\psi}^{\mathfrak{s}}_{k_\zeta, \lfloor \frac{M}{2} \rfloor+1} \right\|_1 \leq \left\| \boldsymbol{\psi}^{\mathfrak{r}}_{k_\zeta, |I_{k_\zeta}^{\mathfrak{rr}}|} \right\|_1 \leq \sum_{i=1}^{|I_{k_\zeta}^{\mathfrak{rr}}|} S_1^{\mathfrak{r}} - (i-1)\overline{\Delta},$$

which implies $S_1^{\mathfrak{s}} < S_1^{\mathfrak{r}}$. Now assume that $L^{\mathfrak{r}} < |I_{k_\zeta}^{\mathfrak{rr}}| \leq \lfloor M/2 \rfloor \leq L^{\mathfrak{s}}$. An ε-stairs-shaped reorganization of $[\mathbf{\Phi}]^{\mathfrak{r}}_{k_\zeta,\cdot}$ exists by virtue of Proposition 6.30. For all $\varepsilon > 0$, the corresponding $\boldsymbol{\beta} \in \mathcal{A}$ and $\mathbf{\Phi}_{k_2,\cdot}(\boldsymbol{\beta})$ also have to abide the suprema from Theorem 6.36. Thus, we can infer

$$\begin{aligned}
&\sum_{i=1}^{\lfloor M/2 \rfloor} S_1^{\mathfrak{s}} - (i-1)\overline{\Delta} \\
&\qquad \leq \left\| \boldsymbol{\psi}^{\mathfrak{r}}_{k_\zeta, |I_{k_\zeta}^{\mathfrak{rr}}|}(\boldsymbol{\alpha}) \right\|_1 = \left\| \boldsymbol{\psi}^{\mathfrak{r}}_{k_2, |I_{k_\zeta}^{\mathfrak{rr}}|}(\boldsymbol{\beta}) \right\|_1 \\
&\qquad \leq \sum_{i=1}^{L^{\mathfrak{r}}} S_1^{\mathfrak{r}} - (i-1)\overline{\Delta} + (|I_{k_\zeta}^{\mathfrak{rr}}| - L^{\mathfrak{r}})\overline{\Delta},
\end{aligned}$$

which implies

$$L^{\mathfrak{r}} S_1^{\mathfrak{s}} + \sum_{i=L^{\mathfrak{r}}+1}^{\lfloor M/2 \rfloor} S_1^{\mathfrak{s}} - (i-1)\overline{\Delta} \leq L^{\mathfrak{r}} S_1^{\mathfrak{r}} + (I_{k_\zeta}^{\mathfrak{rr}}| - L^{\mathfrak{r}})\overline{\Delta}.$$

As $S_1^{\mathfrak{s}} - (i-1)\overline{\Delta} > \overline{\Delta}$ for $i \in \left\{L^{\mathfrak{r}}, \dots, |I_{k_\zeta}^{\mathfrak{rr}}|, \dots, \lfloor M/2 \rfloor\right\}$, we infer

$$\left(\left\lfloor \frac{M}{2} \right\rfloor - L^{\mathfrak{r}}\right)\overline{\Delta} < \left(|I_{k_\zeta}^{\mathfrak{rr}}| - L^{\mathfrak{r}}\right)\overline{\Delta},$$

i.e. $\lfloor M/2 \rfloor < |I_{k_\zeta}^{\mathfrak{rr}}|$, which contradicts $|I_{k_\zeta}^{\mathfrak{rr}}| \leq \lfloor M/2 \rfloor$. Thus, this case is impossible and the other one gives

$$\left\lfloor \frac{M}{2} \right\rfloor \overline{\Delta} \leq S_1^{\mathfrak{s}} < S_1^{\mathfrak{r}}.$$

As we have never exploited the specific value of the sign $\mathfrak{s}$, the analogous reasoning implies $S_1^{\mathfrak{s}} > S_1^{\mathfrak{r}}$, which is a contradiction. We conclude that the assumption (6.17) was false and the claim holds true. □

Algorithm 6.3 Compute $(\boldsymbol{\beta}_{k,\cdot})_{k_1+1\leq k\leq k_2}$ to achieve $[\boldsymbol{\Phi}]^{\mathfrak{s}}_{k_2,a} = [\boldsymbol{\Phi}]^{\mathfrak{s}}_{k_2,b} + \overline{\Delta}$

Require: $\mathfrak{s} \in \{-,+\}$; $\boldsymbol{\Phi}_{k_1,\cdot} \in \mathbb{R}^M$; $a, b \in I^{\mathfrak{s}}_{k_1}$; $[\boldsymbol{\Phi}]^{\mathfrak{s}}_{k_1,a} \geq [\boldsymbol{\Phi}]^{\mathfrak{s}}_{k_1,b}$; $[\boldsymbol{\Phi}]^{\mathfrak{s}}_{k_1,a} + [\boldsymbol{\Phi}]^{\mathfrak{s}}_{k_1,a} \geq \overline{\Delta}$
Require: $0 < \varepsilon < \frac{1}{2}, (\varepsilon \ll \frac{1}{2})$

$k \leftarrow k_1$
while $[\boldsymbol{\Phi}]^{\mathfrak{s}}_{k,a} \neq [\boldsymbol{\Phi}]^{\mathfrak{s}}_{k,b} + \overline{\Delta}$ **do**

$$t^- \leftarrow \begin{cases} 1-\varepsilon, & \text{A}: \quad \boldsymbol{\Phi}_{k,a} - \boldsymbol{\Phi}_{k,b} \leq -2\overline{\Delta}, \\ \frac{\boldsymbol{\Phi}_{k,b} - \boldsymbol{\Phi}_{k,a} - \overline{\Delta}}{2\overline{\Delta}}, & \text{B}: \quad h_{k+1} = \overline{\Delta} \text{ and } \boldsymbol{\Phi}_{k,a} - \boldsymbol{\Phi}_{k,b} \in (-2\overline{\Delta}, -\overline{\Delta}), \\ \frac{\boldsymbol{\Phi}_{k,b} - \boldsymbol{\Phi}_{k,a} + \overline{\Delta}}{2\overline{\Delta}}, & \text{C}: \quad h_{k+1} = \overline{\Delta} \text{ and } \boldsymbol{\Phi}_{k,a} - \boldsymbol{\Phi}_{k,b} \in (-\overline{\Delta}, 0], \\ 1, & \text{D}: \quad \text{else.} \end{cases}$$

$$t^+ \leftarrow \begin{cases} \varepsilon, & \text{A}: \quad \boldsymbol{\Phi}_{k,a} - \boldsymbol{\Phi}_{k,b} \geq 2\overline{\Delta}, \\ \frac{\boldsymbol{\Phi}_{k,b} - \boldsymbol{\Phi}_{k,a} + 3\overline{\Delta}}{2\overline{\Delta}}, & \text{B}: \quad h_{k+1} = \overline{\Delta} \text{ and } \boldsymbol{\Phi}_{k,a} - \boldsymbol{\Phi}_{k,b} \in (\overline{\Delta}, 2\overline{\Delta}), \\ \frac{\boldsymbol{\Phi}_{k,b} - \boldsymbol{\Phi}_{k,a} + \overline{\Delta}}{2\overline{\Delta}}, & \text{C}: \quad h_{k+1} = \overline{\Delta} \text{ and } \boldsymbol{\Phi}_{k,a} - \boldsymbol{\Phi}_{k,b} \in [0, \overline{\Delta}), \\ 1, & \text{D}: \quad \text{else.} \end{cases}$$

$\boldsymbol{\beta}_{k+1,\cdot} \leftarrow t^{\mathfrak{s}} e_a + (1-t^{\mathfrak{s}}) e_b$
$\boldsymbol{\Phi}_{k+1,\cdot} \leftarrow$ (SUR-VC)($\boldsymbol{\Phi}_{k,\cdot}$, $\boldsymbol{\beta}_{k+1,\cdot}$, h_{k+1}, on parity round a if $\mathfrak{s} = -$ else b)
$k \leftarrow k+1$
end while
return $\boldsymbol{\beta}_{k_1+1,\cdot}, \ldots, \boldsymbol{\beta}_{k,\cdot}$

6.5.3 Construction algorithms[5]

Let $\mathfrak{s} \in \{-,+\}$. We start with Algorithm 6.3, which takes a given $[\boldsymbol{\Phi}]^{\mathfrak{s}}_{k_1,\cdot}$ as input. It produces a finite sequence of relaxed controls such that the application of (SUR-VC) along this sequence from k_1 on modifies two entries of $[\boldsymbol{\Phi}]^{\mathfrak{s}}_{k_1,\cdot}$ that sum to a value greater than or equal to $\overline{\Delta}$. After the application, one of these entries takes the value of their average plus $0.5\overline{\Delta}$ and the other one takes the value of their average minus $0.5\overline{\Delta}$.

Lemma 6.38 (Asymptotics and termination of Algorithm 6.3, [82, Lem. 5.1])**.** *Let Assumption 6.27 hold and let the requirements of Algorithm 6.4 be satisfied. Then, Algorithm 6.4 terminates after finitely many iterations such that*

$$\boldsymbol{\Phi}_{k_2,i} = \begin{cases} \boldsymbol{\Phi}_{k_1,i}, & i \notin \{a,b\} \\ \frac{\boldsymbol{\Phi}_{k_1,a} + \boldsymbol{\Phi}_{k_1,b}}{2} + \frac{\overline{\Delta}}{2}, & i = a \text{ and } \mathfrak{s} = + \\ \frac{\boldsymbol{\Phi}_{k_1,a} + \boldsymbol{\Phi}_{k_1,b}}{2} - \frac{\overline{\Delta}}{2}, & i = a \text{ and } \mathfrak{s} = - \\ \frac{\boldsymbol{\Phi}_{k_1,a} + \boldsymbol{\Phi}_{k_1,b}}{2} - \frac{\overline{\Delta}}{2}, & i = b \text{ and } \mathfrak{s} = + \\ \frac{\boldsymbol{\Phi}_{k_1,a} + \boldsymbol{\Phi}_{k_1,b}}{2} + \frac{\overline{\Delta}}{2}, & i = b \text{ and } \mathfrak{s} = - \end{cases}$$

[5] We acknowledge that parts of the results in this subsection, notably Algorithm 6.3 and Lemma 6.38, were developed jointly with Felix Lenders, then at Heidelberg University.

where k_2 denotes the final iteration index. In particular, $\|\boldsymbol{\Phi}_{k_2,\cdot}\|_1 = \|\boldsymbol{\Phi}_{k_1,\cdot}\|_1$.

Proof. We analyze Algorithm 6.3 by investigating the effect of (SUR-VC) in the cases A, B, C, D. Furthermore, we assume inductively that $\|\boldsymbol{\Phi}_{k,\cdot}\|_1 = \|\boldsymbol{\Phi}_{k_1,\cdot}\|_1$, $[\boldsymbol{\Phi}]^{\mathfrak{s}}_{k,a} + [\boldsymbol{\Phi}]^{\mathfrak{s}}_{k,b} \geq \overline{\Delta}$ and $[\boldsymbol{\Phi}]^{\mathfrak{s}}_{k,a} \geq [\boldsymbol{\Phi}]^{\mathfrak{s}}_{k,b} \geq 0$. The algorithm terminates immediately if $[\boldsymbol{\Phi}]^{\mathfrak{s}}_{k,a} - [\boldsymbol{\Phi}]^{\mathfrak{s}}_{k,b} = \overline{\Delta}$.

If case A occurs, we obtain the estimate

$$\boldsymbol{\gamma}_{k+1,a} = \boldsymbol{\Phi}_{k,a} + (1-\varepsilon)h_{k+1} \leq \boldsymbol{\Phi}_{k,b} + (1-\varepsilon)h_{k+1} - 2\overline{\Delta} < \boldsymbol{\Phi}_{k,b} + \varepsilon h_{k+1} = \boldsymbol{\gamma}_{k+1,b}$$

if $\mathfrak{s} = -$ and the estimate

$$\boldsymbol{\gamma}_{k+1,a} = \boldsymbol{\Phi}_{k,a} + \varepsilon h_{k+1} \geq \boldsymbol{\Phi}_{k,b} + \varepsilon h_{k+1} + 2\overline{\Delta} > \boldsymbol{\Phi}_{k,b} + (1-\varepsilon)h_{k+1} = \boldsymbol{\gamma}_{k+1,b}$$

if $\mathfrak{s} = +$. For $\mathfrak{s} = -$, b is the rounding index and for $\mathfrak{s} = +$, a is the rounding index. Regardless of the sign, the application of (SUR-VC) reduces the difference between the entries by

$$[\boldsymbol{\Phi}]^{\mathfrak{s}}_{k,a} - [\boldsymbol{\Phi}]^{\mathfrak{s}}_{k,b} - \left([\boldsymbol{\Phi}]^{\mathfrak{s}}_{k+1,a} - [\boldsymbol{\Phi}]^{\mathfrak{s}}_{k+1,b}\right) = 2(1-\varepsilon)h_{k+1}$$

because $\boldsymbol{\Phi}_{k+1,a} = \boldsymbol{\Phi}_{k,a} + (\varepsilon - 1)h_{k+1}$ and $\boldsymbol{\Phi}_{k+1,b} = \boldsymbol{\Phi}_{k,b} + (1-\varepsilon)h_{k+1}$ for $\mathfrak{s} = +$ and vice versa for $\mathfrak{s} = -$. Regarding well-definedness, we observe that $[\boldsymbol{\Phi}]^{\mathfrak{s}}_{k,b}$ was increased by $(1-\varepsilon)h_{k+1}$ and did not change its sign. Furthermore, $[\boldsymbol{\Phi}]^{\mathfrak{s}}_{k,a}$ was decreased by $(1-\varepsilon)h_{k+1}$, but as the reduction of the gap between the entries is $2(1-\varepsilon)h_{k+1} < 2\overline{\Delta}$, we still have $[\boldsymbol{\Phi}]^{\mathfrak{s}}_{k+1,a} \geq [\boldsymbol{\Phi}]^{\mathfrak{s}}_{k,b}$ yielding $\|\boldsymbol{\Phi}_{k+1,\cdot}\|_1 = \|\boldsymbol{\Phi}_{k,\cdot}\|_1$ and $[\boldsymbol{\Phi}]^{\mathfrak{s}}_{k+1,a} + [\boldsymbol{\Phi}]^{\mathfrak{s}}_{k+1,b} \geq \overline{\Delta}$. Assumption 6.27 guarantees that after finitely many iterations, in which always case A occurs, we reach an iterate with $[\boldsymbol{\Phi}]^{\mathfrak{s}}_{k,a} - [\boldsymbol{\Phi}]^{\mathfrak{s}}_{k,b} \in (0, 2\overline{\Delta})$, i.e. case B, C or D occurs.

If case B occurs, we have $\boldsymbol{\Phi}_{k,b} - \boldsymbol{\Phi}_{k,a} \in (\overline{\Delta}, 2\overline{\Delta})$ if $\mathfrak{s} = -$ and consequently,

$$\frac{\boldsymbol{\Phi}_{k,b} - \boldsymbol{\Phi}_{k,a} - \overline{\Delta}}{2\overline{\Delta}} \in \left(0, \frac{1}{2}\right),$$

which asserts well-definedness for t^-. To assess the effect of (SUR-VC), we compute

$$\boldsymbol{\gamma}_{k,a} = \boldsymbol{\Phi}_{k,a} + \boldsymbol{\beta}_{k+1,a}\overline{\Delta} = \boldsymbol{\Phi}_{k,a} + t^-\overline{\Delta} = \frac{\boldsymbol{\Phi}_{k,a} + \boldsymbol{\Phi}_{k,b}}{2} - \frac{\overline{\Delta}}{2}$$

and

$$\boldsymbol{\gamma}_{k,b} = \boldsymbol{\Phi}_{k,b} + \boldsymbol{\beta}_{k+1,b}\overline{\Delta} = \boldsymbol{\Phi}_{k,b} + (1-t^-)\overline{\Delta} = \frac{\boldsymbol{\Phi}_{k,a} + \boldsymbol{\Phi}_{k,b}}{2} + \frac{3\overline{\Delta}}{2}$$

if $\mathfrak{s} = -$. Furthermore, we have $\boldsymbol{\Phi}_{k,b} - \boldsymbol{\Phi}_{k,a} \in (-2\overline{\Delta}, -\overline{\Delta})$ for $\mathfrak{s} = +$ and consequently,

$$\frac{\boldsymbol{\Phi}_{k,b} - \boldsymbol{\Phi}_{k,a} + 3\overline{\Delta}}{2\overline{\Delta}} \in \left(\frac{1}{2}, 1\right),$$

which asserts well-definedness for t^+. To assess the effect of (SUR-VC), we compute

$$\boldsymbol{\gamma}_{k,a} = \boldsymbol{\Phi}_{k,a} + \boldsymbol{\beta}_{k+1,a}\overline{\Delta} = \boldsymbol{\Phi}_{k,a} + t^+\overline{\Delta} = \frac{\boldsymbol{\Phi}_{k,a} + \boldsymbol{\Phi}_{k,b}}{2} + \frac{3\overline{\Delta}}{2}$$

and

$$\boldsymbol{\gamma}_{k,b} = \boldsymbol{\Phi}_{k,b} + \beta_{k+1,b}\overline{\Delta} = \boldsymbol{\Phi}_{k,b} + (1-t^+)\overline{\Delta} = \frac{\boldsymbol{\Phi}_{k,a} + \boldsymbol{\Phi}_{k,b}}{2} - \frac{\overline{\Delta}}{2}.$$

If case C occurs, we have

$$\frac{\boldsymbol{\Phi}_{k,b} - \boldsymbol{\Phi}_{k,a} + \overline{\Delta}}{2\overline{\Delta}} \in (0,1),$$

which asserts well-definedness for $t^{\mathfrak{s}}$. Furthermore,

$$\boldsymbol{\gamma}_{k,a} = \boldsymbol{\Phi}_{k,a} + \boldsymbol{\beta}_{k+1,a}\overline{\Delta} = \boldsymbol{\Phi}_{k,a} + t^{\mathfrak{s}}\overline{\Delta} = \frac{\boldsymbol{\Phi}_{k,a} + \boldsymbol{\Phi}_{k,b}}{2} + \frac{\overline{\Delta}}{2}$$

and

$$\boldsymbol{\gamma}_{k,b} = \boldsymbol{\Phi}_{k,b} + \beta_{k+1,b}\overline{\Delta} = \boldsymbol{\Phi}_{k,b} + (1-t^{\mathfrak{s}})\overline{\Delta} = \frac{\boldsymbol{\Phi}_{k,a} + \boldsymbol{\Phi}_{k,b}}{2} + \frac{\overline{\Delta}}{2}.$$

In all cases for B and C, the application of (SUR-VC), with the defined behavior in case of non-uniqueness, gives

$$[\boldsymbol{\Phi}]^{\mathfrak{s}}_{k+1,a} = \frac{[\boldsymbol{\Phi}]^{\mathfrak{s}}_{k,a} + [\boldsymbol{\Phi}]^{\mathfrak{s}}_{k,b}}{2} + \frac{\overline{\Delta}}{2} \text{ and } [\boldsymbol{\Phi}]^{\mathfrak{s}}_{k+1,b} = \frac{[\boldsymbol{\Phi}]^{\mathfrak{s}}_{k,a} + [\boldsymbol{\Phi}]^{\mathfrak{s}}_{k,b}}{2} - \frac{\overline{\Delta}}{2}.$$

Thus, we have conserved the properties from the induction hypothesis, i.e. $\|\boldsymbol{\Phi}_{k+1,\cdot}\|_1 = \|\boldsymbol{\Phi}_{k,\cdot}\|_1$, $[\boldsymbol{\Phi}]^{\mathfrak{s}}_{k+1,a} + [\boldsymbol{\Phi}]^{\mathfrak{s}}_{k+1,b} \geq \overline{\Delta}$ and $[\boldsymbol{\Phi}]^{\mathfrak{s}}_{k+1,a} \geq [\boldsymbol{\Phi}]^{\mathfrak{s}}_{k+1,b} \geq 0$ hold. Thus, the termination criterion is satisfied and Algorithm 6.3 terminates with the correct result.

If case D occurs, Assumption 6.27 ensures that it takes a finite number of iterations until case B or C occurs, which then leads to termination with the correct result.

By inspection, we see that we have handled all cases that could possibly occur, or the algorithm would have terminated correctly already, and the proof is finished. □

Remark 6.39 ([82, Rem. 5.2])**.** *Algorithm 6.3 fixes the behavior of* $\arg\max$ *in* (SUR-GEN) *in the case of non-uniqueness. The same reasoning works for other choices if* a *and* b *are updated in every iteration to point to the considered entry with larger (smaller) value. Furthermore, Algorithm 6.5 builds on Algorithm 6.3 and would need additional checks and / or indices to encode increases and decreases of the entries. This incurs additional algorithmic bloat, which we shy away from to keep the construction concise.*

Remark 6.40. *The value of* ε *is not crucial for the asymptotics of Algorithm 6.3, it just has to lie strictly between* 0 *and* 0.5*. The algorithm may need fewer iterations in case A if* ε *is chosen close to* 0*. Furthermore, the algorithm relies on Assumption 6.27. In fact, the construction algorithms are the part in the proof of Theorem 6.25, which require Assumption 6.27. Fortunately, the constructions in Section 6.5.3 serve to obtain a worst-case behavior, which can be assumed to happen at later cells, i.e. in the continuation of the sequences for* $k > N$ *(if one assumes a grid comprising* N *cells). Thus, the bound in Theorem 6.25 for grids consisting of finitely many cells is not affected by this restriction.*

Algorithm 6.4 Compute $(\boldsymbol{\beta}_{k,\cdot})_{k_1+1\leq k\leq k_2}$ to achieve $\boldsymbol{\Phi}_{k_2,b}=0$, $\boldsymbol{\Phi}_{k_2,a}=\boldsymbol{\Phi}_{k_1,a}+\boldsymbol{\Phi}_{k_1,b}$

Require: $\mathfrak{s}\in\{-,+\}$; $\boldsymbol{\Phi}_{k_1,\cdot}\in\mathbb{R}^M$; $a,b\in I_{k_1}^{\mathfrak{s}}$; $[\boldsymbol{\Phi}]_{k_1,a}^{\mathfrak{s}}+[\boldsymbol{\Phi}]_{k_1,b}^{\mathfrak{s}}<\overline{\Delta}$; $[\boldsymbol{\Phi}]_{k_1,a}^{\mathfrak{s}}\geq[\boldsymbol{\Phi}]_{k_1,b}^{\mathfrak{s}}$

1: $k\leftarrow k_1$

2: **while** $[\boldsymbol{\Phi}]_{k,b}^{\mathfrak{s}}>0$ **do**

3: $$t\leftarrow\begin{cases}1, & \text{A}: h_{k+1}<\overline{\Delta},\\ \begin{cases}1-\frac{[\boldsymbol{\Phi}]_{k_1,b}^{+}}{\overline{\Delta}}, & \mathfrak{s}=+\\ \frac{[\boldsymbol{\Phi}]_{k_1,b}^{-}}{\overline{\Delta}}, & \mathfrak{s}=-\end{cases}, & \text{B}: h_{k+1}=\overline{\Delta}.\end{cases}$$

4: $\boldsymbol{\beta}_{k+1,\cdot}\leftarrow e_a(1-t)+e_bt$

5: $\boldsymbol{\Phi}_{k+1,\cdot}\leftarrow$ (SUR-VC)$(\boldsymbol{\Phi}_{k,\cdot},\,\boldsymbol{\beta}_{k+1,\cdot},\,h_{k+1})$

6: $k\leftarrow k+1$

7: **end while**

8: **return** $\boldsymbol{\beta}_{k_1+1,\cdot},\ldots,\boldsymbol{\beta}_{k,\cdot}$

We continue with Algorithm 6.4, which produces a finite sequence of relaxed controls. Applying (SUR-VC) along this sequence modifies two either positive or negative entries of $\boldsymbol{\Phi}_{k_1,\cdot}$ that sum to a value less than $\overline{\Delta}$ such that at the end, one of them takes the value of the sum of both and the other the value zero. The other entries remain unchanged.

Lemma 6.41 (Asymptotics and termination of Algorithm 6.4)**.** *Let Assumption 6.27 hold and let the requirements of Algorithm 6.4 be satisfied. Then, Algorithm 6.4 terminates after finitely many iterations such that*

$$\boldsymbol{\Phi}_{k_2,i}=\begin{cases}\boldsymbol{\Phi}_{k_1,i}, & i\notin\{a,b\}\\ \boldsymbol{\Phi}_{k_1,a}+\boldsymbol{\Phi}_{k_1,b}, & i=a\\ 0, & i=b\end{cases}$$

where k_2 denotes the final iteration index. In particular, $\|\boldsymbol{\Phi}_{k_2,\cdot}\|_1=\|\boldsymbol{\Phi}_{k_1,\cdot}\|_1$.

Proof. Line 4 of Algorithm 6.4 reveals that $\boldsymbol{\beta}_{k+1,i}>0$ is only possible for $i\in\{a,b\}$ for all iterations $k_1\leq k$. Thus, for $i\notin\{a,b\}$, (SUR-VC) ensures the equality $\boldsymbol{\Phi}_{k,i}=\boldsymbol{\Phi}_{k_1,i}$.

In case A, $\boldsymbol{\beta}_{k+1,b}=1$ and $\boldsymbol{\beta}_{k+1,i}=0$ for $i\neq b$, which yields $\boldsymbol{\Phi}_{k+1,\cdot}=\boldsymbol{\Phi}_{k,\cdot}$ by (SUR-VC). In the first iteration in which case B occurs, applying (SUR-VC) gives $\boldsymbol{\Phi}_{k+1,a}=\boldsymbol{\Phi}_{k_1,a}+\boldsymbol{\Phi}_{k_1,b}$ and and $\boldsymbol{\Phi}_{k+1,b}=0$ and Algorithm 6.4 terminates. If $\mathfrak{s}=+$, the rounding occurs in entry b and if $\mathfrak{s}=-$, the rounding occurs in entry a. Assumption 6.27 ensures that case B occurs after finitely many iterations. □

Remark 6.42. *Due to the requirement $[\boldsymbol{\Phi}]_{k_1,a}^{\mathfrak{s}}+[\boldsymbol{\Phi}]_{k_1,b}^{\mathfrak{s}}<\overline{\Delta}$, non-uniqueness of the rounding index cannot occur in the application of* (SUR-VC) *inside Algorithm 6.4.*

We investigate Algorithm 6.5, which builds on Algorithms 6.3 and 6.4 and is the final ingredient to prove Proposition 6.30 and in turn to complete the proof of Theorem 6.25.

Algorithm 6.5 Compute $(\boldsymbol{\beta}_{k,\cdot})_{k_1+1\leq k\leq k_2}$ to achieve $[\boldsymbol{\Phi}]^{\mathfrak{s}}_{k_2,\cdot}$ being ε-stairs-shaped

Require: $\mathfrak{s}\in\{-,+\}$; $\boldsymbol{\Phi}_{k_0,\cdot}\in\mathbb{R}^M$; $J\subset I^{\mathfrak{s}}_{k_0}$; $\varepsilon>0$
Require: j^k_ℓ denotes ℓ-th largest element of $\{[\boldsymbol{\Phi}]^{\mathfrak{s}}_{k,j} : j\in J\}$.
1: $k\leftarrow k_1$, $\kappa\leftarrow k_1$
2: **while** $\exists \ell\in\{1,\dots,|J|-1\} : \left|[\boldsymbol{\Phi}]^{\mathfrak{s}}_{k,j^k_\ell}-[\boldsymbol{\Phi}]^{\mathfrak{s}}_{k,j^k_{\ell+1}}\right|\notin B_\varepsilon(\overline{\Delta})$ and $[\boldsymbol{\Phi}]^{\mathfrak{s}}_{k,j^k_{\ell+1}}\neq 0$ **do**
3: $a,b\leftarrow j^k_1, j^k_1$
4: **for** $\ell=1,\dots,|J|-1$ **do**
5: $a,b\leftarrow\begin{cases} b, j^k_{\ell+1} & \text{if } [\boldsymbol{\Phi}]^{\mathfrak{s}}_{\kappa,b}\geq[\boldsymbol{\Phi}]^{\mathfrak{s}}_{\kappa,j^k_{\ell+1}}\\ j^k_{\ell+1}, b & \text{if } [\boldsymbol{\Phi}]^{\mathfrak{s}}_{\kappa,b}<[\boldsymbol{\Phi}]^{\mathfrak{s}}_{\kappa,j^k_{\ell+1}}\end{cases}$
6: **if** $[\boldsymbol{\Phi}]^{\mathfrak{s}}_{\kappa,a}+[\boldsymbol{\Phi}]^{\mathfrak{s}}_{\kappa,b}\geq\overline{\Delta}$ and $[\boldsymbol{\Phi}]^{\mathfrak{s}}_{\kappa,a}-[\boldsymbol{\Phi}]^{\mathfrak{s}}_{\kappa,b}\neq\overline{\Delta}$ and $[\boldsymbol{\Phi}]^{\mathfrak{s}}_{\kappa,b}\neq 0$ **then**
7: $(\boldsymbol{\beta}_{\kappa+m,\cdot},\boldsymbol{\Phi}_{\kappa+m,\cdot})_{m=1,\dots,L}\leftarrow$ Algorithm 6.3($\mathfrak{s}$, $\boldsymbol{\Phi}_{\kappa,\cdot}$, a, b)
8: **else if** $[\boldsymbol{\Phi}]^{\mathfrak{s}}_{\kappa,a}-[\boldsymbol{\Phi}]^{\mathfrak{s}}_{\kappa,b}\neq\overline{\Delta}$ and $[\boldsymbol{\Phi}]^{\mathfrak{s}}_{\kappa,b}\neq 0$ **then**
9: $(\boldsymbol{\beta}_{\kappa+m,\cdot},\boldsymbol{\Phi}_{\kappa+m,\cdot})_{m=1,\dots,L}\leftarrow$ Algorithm 6.4($\mathfrak{s}$, $\boldsymbol{\Phi}_{\kappa,\cdot}$, a, b)
10: **end if**
11: $\kappa\leftarrow\kappa+L$
12: **end for**
13: $k\leftarrow\kappa$
14: **end while**
15: **return** $\boldsymbol{\beta}_{k_1+1,\cdot},\dots,\boldsymbol{\beta}_{k,\cdot}$

As Algorithms 5.2 and 5.3 from [82] are simplified and merged into Algorithm 6.5, the following Lemma 6.43 comprises arguments from Lemmas 5.3 and 5.4 in [82].

Lemma 6.43 (Asymptotics and termination of Algorithm 6.5)**.** *Let Assumption 6.27 hold and let the requirements of Algorithm 6.5 be satisfied. Then, Algorithm 6.5 terminates after finitely many iterations such that $[\boldsymbol{\Phi}]^{\mathfrak{s}}_{k_2,\cdot}$ is ε-stairs-shaped in J and $\|[\boldsymbol{\Phi}]^{\mathfrak{s}}_{k_2,\cdot}\|_1=\|[\boldsymbol{\Phi}]^{\mathfrak{s}}_{k_1,\cdot}\|_1$ where k_2 denotes the final cell index. Furthermore, $\boldsymbol{\Phi}_{k_1,j}=\ldots=\boldsymbol{\Phi}_{k_2,j}$ for all $j\notin J$.*

Proof. First we argue briefly that the algorithm terminates with the correct result if it terminates. Then, we show that the iterates converge to a point that satisfies the termination criterion after finitely many iterations.

The termination condition in Line 2 ensures that Algorithm 6.5 only terminates if $\boldsymbol{\Phi}_{k,\cdot}$ is ε-stairs-shaped as it just checks Definition 6.29. A quick inspection reveals that entries of $\boldsymbol{\Phi}_{k,\cdot}$ are only modified by Algorithms 6.3 and 6.4. Inductively, this gives $\|\boldsymbol{\Phi}_{k_1,\cdot}\|_1=\|\boldsymbol{\Phi}_{k,\cdot}\|_1$ by virtue of Lemmas 6.38 and 6.41 as well as $\boldsymbol{\Phi}_{k_1,j}=\ldots=\boldsymbol{\Phi}_{k_2,j}$ for $j\notin J$ as a, $b\in J$ holds inductively for all iterations.

Thus, it remains to show termination after finitely many iterations. We notice that by Lines 6 and 8, Algorithm 6.5 does not modify entries that have already been set to zero by the application of Algorithm 6.4. Consequently, this can only happen finitely

many times as the algorithm would terminate after the iteration, in which all entries have been set to zero. We conclude that after finitely many iterations the set of indices of the entries that are modified in the **for**-loop by the application of Algorithm 6.3 does not change anymore. By Line 6, the application of Algorithm 6.3 is always well-defined, i.e. it satisfies the requirements of Algorithm 6.3 and Lemma 6.38. Thus, we restrict ourselves to the case where Line 6 is never executed and the entries that have the value zero can safely be ignored in the further considerations without any loss of generality.

Next, we show that one **for**-loop produces an update on the values in $\mathbf{\Phi}_{k,\cdot}$ that is equivalent to the application of a linear transformation. Then, a spectral analysis of the transformation yields the desired convergence.

We start at cell index k with arbitrary value $[\mathbf{\Phi}]^{\mathfrak{s}}_{k,\cdot}$. To keep track of the considered entries of $[\mathbf{\Phi}]^{\mathfrak{s}}_{\kappa,\cdot}$ during the iterations of the **for**-loop, we denote the cell index after the ℓ-th iteration of the **for**-loop by κ_ℓ and the index before the **for**-loop, or alternatively, at the end of the previous **for**-loop, by $\kappa_0 = k$. Furthermore, we refer to the cell index b in the ℓ-th iteration of the **for**-loop by b_ℓ.

The first iteration of the **for**-loop executes Algorithm 6.3, which yields

$$[\mathbf{\Phi}]^{\mathfrak{s}}_{\kappa_1,j_1^k} = \frac{[\mathbf{\Phi}]^{\mathfrak{s}}_{k,j_1^k} + [\mathbf{\Phi}]^{\mathfrak{s}}_{k,j_2^k} + \overline{\Delta}}{2},$$
$$[\mathbf{\Phi}]^{\mathfrak{s}}_{\kappa_1,j_2^k} = \frac{[\mathbf{\Phi}]^{\mathfrak{s}}_{k,j_2^k} + [\mathbf{\Phi}]^{\mathfrak{s}}_{k,j_2^k} - \overline{\Delta}}{2}$$

by Lemma 6.38 and leaves the other entries unchanged. As the j_1^k-th entry of $[\mathbf{\Phi}]^{\mathfrak{s}}_{\kappa_1,\cdot}$ has been increased, we get $j_1^k = j_1^{\kappa_1}$, but we do not know the rank of $[\mathbf{\Phi}]^{\mathfrak{s}}_{\kappa_1,j_2^k}$ in the ordering by value anymore as it has been decreased. The entry is named b in Algorithm 6.5 and with the notation introduced above, we have $b_1 = j_2^n$. For the second iteration of the **for**-loop, we notice that either the entry at index b_1 that has just been decreased or the entry at index j_3^k is now the entry $j_2^{\kappa_1}$. Furthermore, $b_1 \neq j_3^k$. Thus, the second application of Algorithm 6.3 gives

$$[\mathbf{\Phi}]^{\mathfrak{s}}_{\kappa_2,j_2^{\kappa_2}} = \frac{[\mathbf{\Phi}]^{\mathfrak{s}}_{\kappa_1,b_1} + [\mathbf{\Phi}]^{\mathfrak{s}}_{k,j_3^k} + \overline{\Delta}}{2},$$
$$[\mathbf{\Phi}]^{\mathfrak{s}}_{\kappa_2,b_2} = \frac{[\mathbf{\Phi}]^{\mathfrak{s}}_{\kappa_1,b_1} + [\mathbf{\Phi}]^{\mathfrak{s}}_{k,j_3^k} - \overline{\Delta}}{2}$$

by Algorithm 6.5 and the fact that $[\mathbf{\Phi}]^{\mathfrak{s}}_{\kappa_1,b_1} < [\mathbf{\Phi}]^{\mathfrak{s}}_{k,j_1^k}$ and $[\mathbf{\Phi}]^{\mathfrak{s}}_{\kappa_1,j_3^k} \leq [\mathbf{\Phi}]^{\mathfrak{s}}_{k,j_2^k}$. This also implies $j_1^{\kappa_2} = j_1^{\kappa_1}$. Continuing this argument inductively over the iterations of the **for**-loop gives $j_\ell^{\kappa_{\ell-1}} \in \{j_{\ell+1}^k, b_{\ell-1}\}$, $j_{\ell+1}^k \neq b_{\ell-1}$ and the application of Algorithm 6.3 gives

$$[\mathbf{\Phi}]^{\mathfrak{s}}_{\kappa_\ell,j_\ell^{\kappa_\ell}} = \frac{[\mathbf{\Phi}]^{\mathfrak{s}}_{\kappa_{\ell-1},b_{\ell-1}} + [\mathbf{\Phi}]^{\mathfrak{s}}_{k,j_\ell^k} + \overline{\Delta}}{2},$$
$$[\mathbf{\Phi}]^{\mathfrak{s}}_{\kappa_\ell,b_\ell} = \frac{[\mathbf{\Phi}]^{\mathfrak{s}}_{\kappa_{\ell-1},b_{\ell-1}} + [\mathbf{\Phi}]^{\mathfrak{s}}_{k,j_\ell^k} - \overline{\Delta}}{2}.$$

Furthermore, we have $j_i^{\kappa_\ell} = j_i^{\kappa_{\ell-1}} = \ldots = j_i^{\kappa_i}$ for all $i \in \{1, \ldots, \ell-1\}$ and $\ell \in \{2, \ldots, |J|-1\}$. Thus, in the ℓ-th iteration of the **for**-loop, the entry $[\mathbf{\Phi}]^{\mathfrak{s}}_{\kappa_\ell, j_\ell^{\kappa_{|J|-1}}} = \ldots = [\mathbf{\Phi}]^{\mathfrak{s}}_{\kappa_{|J|-1}, j_\ell^{\kappa_{|J|-1}}}$ is computed and assigned. In the last iteration, also the smallest entry $\mathbf{\Phi}_{\kappa_{|J|-1}, j_{|J|}^{\kappa_{|J|-1}}}$ is computed and assigned, i.e. $b_{\kappa_{|J|-1}} = j_{|J|}^{\kappa_{|J|-1}}$. From the recursive formulae just derived, it is immediate that the entries in $[\mathbf{\Phi}]^{\mathfrak{s}}_{\kappa_\ell, \cdot}$ depend linearly (or affinely) on the entries in $[\mathbf{\Phi}]^{\mathfrak{s}}_{\kappa_0, \cdot}$. Adding an extra line for $\overline{\Delta}$, we can cast the recursive formulae into the update matrix below that represents the effect of one run of the for-loop, or alternatively, one iteration of the **while**-loop.

$$\begin{pmatrix} \overline{\Delta} \\ [\mathbf{\Phi}]^{\mathfrak{s}}_{\kappa_{|J|-1}, j_1^{\kappa_{|J|-1}}} \\ [\mathbf{\Phi}]^{\mathfrak{s}}_{\kappa_{|J|-1}, j_2^{\kappa_{|J|-1}}} \\ [\mathbf{\Phi}]^{\mathfrak{s}}_{\kappa_{|J|-1}, j_3^{\kappa_{|J|-1}}} \\ \vdots \\ [\mathbf{\Phi}]^{\mathfrak{s}}_{\kappa_{|J|-1}, j_{|J|-1}^{\kappa_{|J|-1}}} \\ [\mathbf{\Phi}]^{\mathfrak{s}}_{\kappa_{|J|-1}, j_{|J|}^{\kappa_{|J|-1}}} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & \ldots & 0 \\ \frac{1}{2} & \frac{1}{2} & \frac{1}{2} & 0 & 0 & \ldots & 0 \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{2} & 0 & \ldots & 0 \\ \frac{1}{8} & \frac{1}{8} & \frac{1}{8} & \frac{1}{4} & \frac{1}{2} & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \ddots & \ddots & \vdots \\ \frac{1}{2^{|J|-1}} & \frac{1}{2^{|J|-1}} & \frac{1}{2^{|J|-1}} & \frac{1}{2^{|J|-2}} & \frac{1}{2^{|J|-3}} & \ldots & \frac{1}{2} \\ \frac{1-2^{|J|-1}}{2^{|J|-1}} & \frac{1}{2^{|J|-1}} & \frac{1}{2^{|J|-1}} & \frac{1}{2^{|J|-2}} & \frac{1}{2^{|J|-3}} & \ldots & \frac{1}{2} \end{pmatrix} \begin{pmatrix} \overline{\Delta} \\ \mathbf{\Phi}_{\kappa_0, j_1^{\kappa_0}} \\ \mathbf{\Phi}_{\kappa_0, j_2^{\kappa_0}} \\ \mathbf{\Phi}_{\kappa_0, j_3^{\kappa_0}} \\ \vdots \\ \mathbf{\Phi}_{\kappa_0, j_{|J|-1}^{\kappa_0}} \\ \mathbf{\Phi}_{\kappa_0, j_{|J|}^{\kappa_0}} \end{pmatrix}$$

To establish a convergence result, we change our vantage point to the difference between two subsequent entries (in the ordering by value) of $[\mathbf{\Phi}]^{\mathfrak{s}}_{\kappa_{|J|-1}, i}$ for $i \in J$. More specifically, we consider

$$d_\ell^k \coloneqq [\mathbf{\Phi}]^{\mathfrak{s}}_{k, j_\ell^k} - [\mathbf{\Phi}]^{\mathfrak{s}}_{k, j_{\ell+1}^k} \geq 0$$

for $\ell \in \{1, \ldots, |J|-1\}$. We deduce

$$d_{|J|-1}^{\kappa_{|J|-1}} = \overline{\Delta},$$

which follows from the linear transformation above and also from the fact that the **for**-loop finishes with an application of Algorithm 6.3 that sets the difference between the two smallest entries under consideration to $\overline{\Delta}$. For the change from d^{κ_0} to $d^{\kappa_{|J|-1}}$ due to one run of the **for**-loop, we obtain

$$d_\ell^{\kappa_{|J|-1}} = \frac{\overline{\Delta} + \sum_{i=1}^{\ell+1} d_i^{\kappa_0} 2^{i-1}}{2^{\ell+1}}$$

for $\ell \in \{1, \ldots, |J|-1\}$, see Lemma A.2 for the details. We combine the update of the d_i^k with the largest entry of $[\mathbf{\Phi}]^{\mathfrak{s}}_{k,i}$, $i \in J$ into the linear update formula

$$\begin{pmatrix} [\mathbf{\Phi}]^{\mathfrak{s}}_{\kappa_{|J|-1}, j_1^{\kappa_{|J|-1}}} \\ d_1^{\kappa_{|J|-1}} \\ d_2^{\kappa_{|J|-1}} \\ \vdots \\ d_{|J|-2}^{\kappa_{|J|-1}} \\ d_{|J|-1}^{\kappa_{|J|-1}} \end{pmatrix} = \begin{pmatrix} 1 & -\frac{1}{2} & 0 & 0 & \ldots & \frac{1}{2} \\ 0 & \frac{1}{4} & \frac{1}{2} & 0 & \ldots & \frac{1}{4} \\ 0 & \frac{1}{8} & \frac{1}{4} & \frac{1}{2} & \ldots & \frac{1}{8} \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \frac{1}{2^{|J|-1}} & \frac{1}{2^{|J|-2}} & \frac{1}{2^{|J|-3}} & \ldots & \frac{1}{2} + \frac{1}{2^{|J|-1}} \\ 0 & 0 & 0 & 0 & \ldots & 1 \end{pmatrix} \begin{pmatrix} [\mathbf{\Phi}]^{\mathfrak{s}}_{\kappa_0, j_1^{\kappa_0}} \\ d_1^{\kappa_0} \\ d_2^{\kappa_0} \\ \vdots \\ d_{|J|-2}^{\kappa_0} \\ d_{|J|-1}^{\kappa_0} \end{pmatrix}.$$

For the remainder of this proof, we denote the update matrix above by the symbol T. Considering its first line and first column, we obtain that e_1 is an eigenvector to the eigenvalue 1. Furthermore, 1 bounds the other eigenvalues from above by the Gershgorin circle theorem. The minor from second column and row on is a row-stochastic matrix. Thus, $\begin{pmatrix} 0 & 1 & \dots & 1 \end{pmatrix}^T$ is an eigenvector of the eigenvalue 1. The last row reveals $d_{|J|-1}^{\kappa_{|J|-1}} = d_{|J|-1}^{\kappa_0}$ and the first row yields $v_2 = v_{|J|}$ for every eigenvector v of the eigenvalue 1. Inductively, we obtain $v_2 = v_3 = v_4 = \ldots = v_{|J|}$. We obtain a geometric multiplicity of 2 of the eigenvalue 1 with the corresponding eigenspace

$$\text{span} \left\{ \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ \vdots \\ 1 \end{pmatrix} \right\}.$$

The matrix also maps coordinate-wise nonnegative vectors to coordinate-wise nonnegative vectors. Thus, analogously to the convergence of the von-Mises-iteration, we get convergence to an element in the eigenspace to the eigenvalue 1. The last row of T gives $d_{|J|-1}^{\kappa_{|J|-1}} = \overline{\Delta}$, which yields

$$\begin{pmatrix} d_1^k \\ \vdots \\ d_{|J|-1}^k \end{pmatrix} \underset{k\to\infty}{\to} \begin{pmatrix} d_{|J|-1}^{\kappa_{|J|-1}} \\ \vdots \\ d_{|J|-1}^{\kappa_{|J|-1}} \end{pmatrix} = \begin{pmatrix} \overline{\Delta} \\ \vdots \\ \overline{\Delta} \end{pmatrix}.$$

The convergence of $[\mathbf{\Phi}]_{k,j_1^k}^{\mathbf{5}}$ follows. Due to the limit behavior just obtained, $\mathbf{\Phi}_{k,\cdot}$ is ε-stairs-shaped in J after finitely many iterations, which closes the proof. □

Lemma 6.43 proves Proposition 6.30, the final step for Theorem 6.25.

6.6 The integrality gap for (SUR-GEN)

There is no good bound. Consider the specialization of (SUR-GEN) with $F_k = \{1\}$ for all $k \in \{1, \ldots, N\}$. Let $\boldsymbol{\alpha}$ be a relaxed control with $\boldsymbol{\alpha}_1 = 0$ on $\overline{\Omega}$. Then, $\boldsymbol{\phi}_1 = -1$ on $\overline{\Omega}$ and

$$d(\boldsymbol{\omega}, \boldsymbol{\alpha}) = \sup_k \left\| \int_{\bigcup_{k'=1}^k \mathcal{T}_{k'}} \begin{pmatrix} -1 \\ \boldsymbol{\alpha}_2(x) \\ \cdots \\ \boldsymbol{\alpha}_M(x) \end{pmatrix} \mathrm{d}x \right\|_\infty = \lambda(\Omega)$$

which is the worst possible outcome and independent of the rounding grid.

Chapter 7

Relaxed control approximation

This chapter serves to achieve contribution (b) in (1.1), i.e. we prove

$$d^{(n)}(\boldsymbol{\omega}^{(n)}, \boldsymbol{\alpha}) \to 0 \quad \underset{(b)}{\Rightarrow} \quad \boldsymbol{\omega}^{(n)} \rightharpoonup \boldsymbol{\alpha}$$

for a sequence of rounding grids with corresponding integrality gaps $(d^{(n)})_n$, a binary control sequence $(\boldsymbol{\omega}^{(n)})_n$ and a relaxed control $\boldsymbol{\alpha}$. We begin with the case of a one-dimensional domain and $d_{\mathrm{1D}}(\boldsymbol{\omega}^{(n)}, \boldsymbol{\alpha}) \to 0$ and continue with the more general multi-dimensional case. The first case, proven in Theorem 7.3, is straightforward as it depends less on the rounding algorithm. To deduce (b) from the results in Chapter 6 in the multi-dimensional case in Theorem 7.9, we require additional assumptions on the sequence of rounding grids, on which the rounding algorithm is executed.

7.1 The one-dimensional case

We consider $d_{\mathrm{1D}}(\boldsymbol{\omega}^{(n)}, \boldsymbol{\alpha}) \to 0$, which is ensured for (SUR) and (SUR-VC) with the findings from Chapter 6 if the rounding grids, used to compute the $\boldsymbol{\omega}^{(n)}$, consist of subsequent intervals, see Proposition 6.2. The claim follows from the integration by parts formula. As the latter is not directly applicable, we provide an additional approximation argument in the following lemma. This argument is a key step to drop the differentiability assumptions imposed in [51, 68, 100] for the convergence analysis of the resulting state vector sequences. This section follows the arguments of the author in [80, Sect 2.1].

Lemma 7.1 ([80, Lem. 2.1])**.** *Let X be a Banach space. Let $f_i \in L^1((0,T), X)$ for $i \in \{1, \dots, M\}$. Let $(\boldsymbol{\phi}^{(n)})_n \subset L^\infty((0,T), \mathbb{R}^M)$ be bounded and of vanishing integrality gap, i.e. $\|\boldsymbol{\Phi}_{\mathrm{1D}}^{(n)}\|_{L^\infty} \to 0$. Let $\varepsilon > 0$. Then, we obtain the convergence*

$$\sup_{t \in [0,T]} \left\| \int_0^t \sum_{i=1}^M \boldsymbol{\phi}_i^{(n)}(s) f_i(s) \, \mathrm{d}s \right\|_X \to 0.$$

Proof. First, it suffices to consider the case $M = 1$, which removes the sum. We define $C_{\boldsymbol{\phi}} \coloneqq \sup_{n \in \mathbb{N}} \|\boldsymbol{\phi}^{(n)}\|_{L^\infty}$ and as $(\boldsymbol{\phi}^{(n)})_n$ is bounded, we have $C_{\boldsymbol{\phi}} < \infty$.

Let $\varepsilon > 0$. Proposition B.29 gives $\overline{C^\infty([0,T],X)}^{\|\cdot\|_{L^1}} = L^1((0,T),X)$ and thus, there exists $g \in C^\infty([0,T],X)$ with $C_g \coloneqq \|g\|_{L^\infty((0,T),X)} + T\|g'\|_{L^\infty((0,T),X)}$ such that

$$\|f - g\|_{L^1((0,T),X)} < \frac{\varepsilon}{2C_{\boldsymbol{\phi}}}.$$

We insert an auxiliary zero into the term of interest

$$\int_0^t f(s)\boldsymbol{\phi}^{(n)}(s)\,\mathrm{d}s = \int_0^t g(s)\boldsymbol{\phi}^{(n)}(s)\,\mathrm{d}s + \int_0^t \boldsymbol{\phi}^{(n)}(s)(f(s) - g(s))\,\mathrm{d}s$$

and apply the integration by parts formula to the first summand. This yields

$$\int_0^t g(s)\boldsymbol{\phi}^{(n)}(s)\,\mathrm{d}s = g(t)\boldsymbol{\Phi}_{\mathsf{1D}}^{(n)}(t) - \int_0^t g'(s)\boldsymbol{\Phi}_{\mathsf{1D}}^{(n)}(s)\,\mathrm{d}s,$$

which can be bounded with the estimate

$$\left\|\int_0^t g(s)\boldsymbol{\phi}^{(n)}(s)\,\mathrm{d}s\right\|_X \leq C_g\|\boldsymbol{\Phi}_{\mathsf{1D}}^{(n)}\|_{L^\infty}.$$

By assumption, $\boldsymbol{\Phi}_{\mathsf{1D}}^{(n)} \to 0$ uniformly, which asserts the existence of $n_0 \in \mathbb{N}$ such that for all $n \geq n_0$, the estimate

$$\|\boldsymbol{\Phi}_{\mathsf{1D}}^{(n)}\|_{L^\infty} < \frac{\varepsilon}{2C_g}$$

holds true. We put the estimates together and obtain

$$\begin{aligned}\sup_{t\in[0,T]}\left\|\int_0^t f(s)\boldsymbol{\phi}^{(n)}(s)\,\mathrm{d}s\right\|_X &\leq \sup_{t\in[0,T]}\left\|\int_0^t g(s)\boldsymbol{\phi}^{(n)}(s)\,\mathrm{d}s\right\|_X + C_{\boldsymbol{\phi}}\|f-g\|_{L^1((0,T),X)}\\ &< C_g\frac{\varepsilon}{2C_g} + C_{\boldsymbol{\phi}}\frac{\varepsilon}{2C_{\boldsymbol{\phi}}} = \varepsilon\end{aligned}$$

for all $n \geq n_0$. □

We give an example, also originating from [80], to illustrate that the result works indeed for $f_i \in L^1((0,T))\backslash W^{1,1}((0,T))$.

Example 7.2 ([80, Example 2.2]). *We define the following Weierstraß function*

$$\begin{aligned}f &: [0,2\pi] \to \mathbb{R},\\ f(x) &\coloneqq \lim f^{(n)}(x),\\ f^{(n)}(x) &\coloneqq \sum_{k=0}^{n-1}\frac{2^k\sin(2^k x)}{3^k},\end{aligned}$$

which is nowhere differentiable. We also define the sequence $(\boldsymbol{\phi}^{(n)})_n$ with the help of an equidistant grid with cell volumes $\frac{2\pi}{2^n}$ as

$$\begin{aligned}\boldsymbol{\phi}^{(n)} &: [0,2\pi] \to [-1,1],\\ \boldsymbol{\phi}^{(n)}(x) &\coloneqq \begin{cases} 1 & : \ x \in 2\pi\cdot\left[\frac{2i}{2^n}, \frac{2i+1}{2^n}\right] & i \in \{0,\dots,2^{n-1}-1\},\\ -1 & : \ x \in 2\pi\cdot\left[\frac{2i+1}{2^n}, \frac{2i+2}{2^n}\right] & i \in \{0,\dots,2^{n-1}-1\}.\end{cases}\end{aligned}$$

The definition of the $\boldsymbol{\phi}^{(n)}$ asserts

$$\int_0^{2\pi} \frac{2^k \sin(2^k x)}{3^k} \boldsymbol{\phi}^{(n)}(x)\,\mathrm{d}x = \frac{2^k}{3^k} \begin{cases} \int_0^{2\pi} |\sin(2^k x)|\,\mathrm{d}x & : \quad k+1 = n \\ 0 & : \quad k+1 \neq n \end{cases}.$$

We note that the sin *terms oscillate inside the intervals where the $f^{(n)}$ terms are constant and cancel each other for $k \geq n$. In case $k \leq n-2$, the $f^{(n)}$ oscillate and cancel themselves within intervals where the* sin *terms have constant sign and are symmetric with respect to their extreme point in this segment. We apply the Lebesgue dominated convergence theorem and arrive at*

$$\begin{aligned} \int_0^{2\pi} f(x)\boldsymbol{\phi}^{(n)}(x)\,\mathrm{d}x &= \lim_{m\to\infty} \int_0^{2\pi} \sum_{k=0}^{m} \frac{2^k \sin(2^k x)}{3^k} \boldsymbol{\phi}^{(n)}(x)\,\mathrm{d}x \\ &= \frac{2^{n-1}}{3^{n-1}} \int_0^{2\pi} |\sin(2^{n-1}x)|\,\mathrm{d}x \leq \frac{2^{n-1}}{3^{n-1}} 2\pi \underset{n\to\infty}{\to} 0. \end{aligned}$$

Finally, we note that the $\boldsymbol{\phi}^{(n)}$ are chosen to provide an instructive argument, which is not that simple for arbitrary sequences $(\boldsymbol{\phi}^{(n)})_n$ of vanishing integrality gap.

Lemma 7.1 yields a characterization of the convergence of $(\boldsymbol{\omega}^{(n)})_n$ and $(\boldsymbol{\phi}^{(n)})_n$ in weak topologies of L^p-spaces, which we summarize in Theorem 7.3.

Theorem 7.3 ([80, Thm 2.4])**.** *Let $\Omega_T = (0,T)$. Let $\boldsymbol{\alpha}$ be a relaxed control. Let $(\boldsymbol{\omega}^{(n)})_n$ be a sequence of binary controls. Let $d_{\mathrm{1D}}(\boldsymbol{\omega}^{(n)}, \boldsymbol{\alpha}) \to 0$. Then,*

$$\boldsymbol{\omega}^{(n)} \rightharpoonup \boldsymbol{\alpha} \text{ in } L^p((0,T),\mathbb{R}^M) \text{ for } 1 \leq p < \infty$$

and

$$\boldsymbol{\omega}^{(n)} \rightharpoonup^* \boldsymbol{\alpha} \text{ in } L^p((0,T),\mathbb{R}^M) \text{ for } 1 < p \leq \infty.$$

Proof. The prerequisites enable us to use Lemma 7.1 with the choice $X = \mathbb{R}^M$ yielding $\boldsymbol{\phi}^{(n)} \rightharpoonup^* 0$ in $L^\infty((0,T),\mathbb{R}^M)$, which is equivalent to $\boldsymbol{\omega}^{(n)} \rightharpoonup^* \boldsymbol{\alpha}$ in $L^\infty((0,T),\mathbb{R}^M)$. The assertion of the other claims is immediate as the test function space $L^1((0,T),\mathbb{R}^M)$ is the largest $L^p((0,T),\mathbb{R}^M)$ space for $p \geq 1$. □

7.2 The multi-dimensional case

This section establishes the desired weak approximation of a relaxed control $\boldsymbol{\alpha}$ by a sequence of binary controls $\boldsymbol{\omega}^{(n)}$, i.e. $\boldsymbol{\omega}^{(n)} \rightharpoonup \boldsymbol{\alpha}$, for $d^{(n)}(\boldsymbol{\omega}^{(n)}, \boldsymbol{\alpha}) \to 0$ in the multi-dimensional case. The proof of the one-dimensional case from Section 7.1 cannot be applied as we do not have a suitable analog to the integration by parts formula available. We impose additional assumptions on the sequence of rounding grids to compute the $\boldsymbol{\omega}^{(n)}$. This section follows the arguments of the author in [79, Sect. 4].

We begin by bounding the duality pairing of a control deviation and an L^1 function.

Lemma 7.4 ([79, Lem. 4.3]). *Let $\{\mathcal{T}_1, \dots, \mathcal{T}_N\}$ be a rounding grid for a bounded domain $\Omega_T \subset \mathbb{R}^d$. Let $\boldsymbol{\alpha}$ and $\boldsymbol{\omega}$ be relaxed controls, which are a.e. constant on each cell $\mathcal{T}_i$. Let $\boldsymbol{\phi} \coloneqq \boldsymbol{\alpha} - \boldsymbol{\omega}$. Then, for all $i \in \{1, \dots, M\}$ and all $f \in L^1(\Omega_T)$, we can estimate*

$$\left| \langle \boldsymbol{\phi}_i, f \rangle_{L^\infty, L^1} \right| \leq 2d(\boldsymbol{\omega}, \boldsymbol{\alpha}) \|f\|_{L^1}.$$

Proof. Let $i \in \{1, \dots, M\}$ be fixed. We set forth to bound the term $\left| \int_{\Omega_T} \boldsymbol{\phi}_i f \right|$ and note

$$\left| \int_{\Omega_T} \boldsymbol{\phi}_i(s) f(s) \, \mathrm{d}s \right| = \left| \sum_{k=1}^{N} \int_{\mathcal{T}_k} \boldsymbol{\phi}_i(s) f(s) \, \mathrm{d}s \right|.$$

As $\boldsymbol{\alpha}$ and $\boldsymbol{\omega}$ are constant per grid cell, so is $\boldsymbol{\phi}$. This yields

$$\left| \int_{\Omega_T} \boldsymbol{\phi}_i(s) f \, \mathrm{d}s \right| = \left| \sum_{j=1}^{N} \bar{\boldsymbol{\phi}}_{ij} \bar{f}_j \right|$$

for the definitions $\bar{\boldsymbol{\phi}}_{ik} \coloneqq \frac{1}{\lambda(\mathcal{T}_k)} \int_{\mathcal{T}_k} \boldsymbol{\phi}_i$ and $\bar{f}_k \coloneqq \int_{\mathcal{T}_k} f$.

For all $n \in \{1, \dots, N\}$, the definition of d, see Definition 6.1, gives

$$-d(\boldsymbol{\omega}, \boldsymbol{\alpha}) \leq \sum_{k=1}^{n} \bar{\boldsymbol{\phi}}_{ik} \leq d(\boldsymbol{\omega}, \boldsymbol{\alpha}). \tag{7.1}$$

Due to the uniform boundedness of the partial sums in d, we can bound

$$-2d(\boldsymbol{\omega}, \boldsymbol{\alpha}) \leq \bar{\boldsymbol{\phi}}_{ik} \leq 2d(\boldsymbol{\omega}, \boldsymbol{\alpha}).$$

for all $k \in \{1, \dots, N\}$. In particular, the estimates

$$\bar{\boldsymbol{\phi}}_{in} \underset{(7.1)}{\leq} d(\boldsymbol{\omega}, \boldsymbol{\alpha}) - \sum_{k=1}^{n-1} \bar{\boldsymbol{\phi}}_{ik} \underset{(7.1)}{\leq} 2d(\boldsymbol{\omega}, \boldsymbol{\alpha})$$

and

$$\bar{\boldsymbol{\phi}}_{in} \underset{(7.1)}{\geq} -d(\boldsymbol{\omega}, \boldsymbol{\alpha}) - \sum_{k=1}^{n-1} \bar{\boldsymbol{\phi}}_{ik} \underset{(7.1)}{\geq} -2d(\boldsymbol{\omega}, \boldsymbol{\alpha})$$

hold for $n \geq 2$ and for the case $n = 1$, the estimate holds by (7.1). Combining these two considerations, we conclude the proof with the estimate

$$\begin{aligned} \left| \sum_{k=1}^{N} \bar{\boldsymbol{\phi}}_{ik} \bar{f}_j \, \mathrm{d}s \right| &\leq \sum_{k=1}^{N} |\bar{\boldsymbol{\phi}}_{ik}| |\bar{f}_k| \\ &\leq 2d(\boldsymbol{\omega}, \boldsymbol{\alpha}) \sum_{i=1}^{N} |\bar{f}_j| \leq 2d(\boldsymbol{\omega}, \boldsymbol{\alpha}) \int_{\Omega_T} |f(s)| \, \mathrm{d}s = 2d(\boldsymbol{\omega}, \boldsymbol{\alpha}) \|f\|_{L^1}. \end{aligned}$$

□

Remark 7.5. *The result of Lemma 7.4 only depends on the order of the grid cells during rounding through $d(\boldsymbol{\omega}, \boldsymbol{\alpha})$. If* (SUR) *or* (SUR-VC) *is employed for rounding, the considerations in Chapter 6 establish bounds on $d(\boldsymbol{\omega}, \boldsymbol{\alpha})$ that hold uniformly for all permutations of the grid cells.*

We provide a sufficient condition on grid refinements to deduce $\boldsymbol{\omega}^{(n)} \rightharpoonup \boldsymbol{\alpha}$.

Definition 7.6 (Admissible sequences of refined rounding grids, [79, Def. 4.4])**.** *Let $\Omega_T \subset \mathbb{R}^d$ be a bounded domain. A sequence of rounding grids $\left(\left\{\mathcal{T}_1^{(n)}, \dots, \mathcal{T}_{N^{(n)}}^{(n)}\right\}\right)_n$ with corresponding sequence of mesh sizes $(\overline{\Delta}^{(n)})_n$ is called an* ***admissible sequence of refined rounding grids of*** *Ω_T if*

1. *$\overline{\Delta}^{(n)} \to 0$,*
2. *for all $n \in \mathbb{N}$ and all $k \in \{1, \dots, N^{(n+1)}\}$, there exists $\ell \in \{1, \dots, N^{(n)}\}$ such that $\mathcal{T}_k^{(n+1)} \subset \mathcal{T}_\ell^{(n)}$,*
3. *the cells* ***shrink regularly****, i.e. there exists $C > 0$ such that for each $\mathcal{T}_k^{(n)}$, there exists a ball $B_k^{(n)}$ such that $\mathcal{T}_j^{(n)} \subset B_k^{(n)}$ and $\lambda(\mathcal{T}_k^{(n)}) \geq C\lambda(B_k^{(n)})$.*

Remark 7.7 ([79, Rem. 4.5])**.** *Definition 7.6 is related to refinements of finite element triangulations, namely to demanding a quasi-uniform mesh, which is refined with an isotropic strategy, see* [5]*. For our purposes, it suffices to restrict the eccentricity of the cells with a bound on the ratio between the volumes of a cell and the circumscribed sphere and we do not need to restrict the ratio between the diameters of the cell and an inscribed sphere.*

We combine Lemma 7.4 and Definition 7.6 with an approximation argument to deduce $\boldsymbol{\omega}^{(n)} \rightharpoonup^* \boldsymbol{\alpha}$.

Lemma 7.8 ([79, Lem. 4.6])**.** *Let $\left(\left\{\mathcal{T}_1^{(n)}, \dots, \mathcal{T}_{N^{(n)}}^{(n)}\right\}\right)_n$ be an admissible sequence of refined rounding grids of a bounded domain $\Omega_T \subset \mathbb{R}^d$ with corresponding integrality gaps $(d^{(n)})_n$. Let $\boldsymbol{\alpha}$ be a relaxed control and for all $n \in \mathbb{N}$, let $\boldsymbol{\omega}^{(n)}$ be a binary control that is a.e. constant on the grid cells $\mathcal{T}_1^{(n)}, \dots, \mathcal{T}_{N^{(n)}}^{(n)}$. Furthermore, we assume*

$$d^{(n)}(\boldsymbol{\omega}^{(n)}, \boldsymbol{\alpha}) \to 0. \tag{7.2}$$

Let $\boldsymbol{\phi}^{(n)} \coloneqq \boldsymbol{\alpha} - \boldsymbol{\omega}^{(n)}$. Then, the convergence

$$\left\langle \boldsymbol{\phi}_i^{(n)}, f \right\rangle_{L^\infty, L^1} \to 0$$

holds for all $i \in \{1, \dots, M\}$ and all $f \in L^1(\Omega_T)$.

Proof. Let $i \in \{1, \dots, M\}$ and $f \in L^1(\Omega_T)$ be fixed. We begin with an approximation of $\boldsymbol{\alpha}$, which serves to leverage Lemma 7.4.

As the cells $\mathcal{T}_k^{(n)}$ partition Ω_T for all n, we can define the functions $\widetilde{\boldsymbol{\alpha}}^{(n)}$ as

$$\widetilde{\boldsymbol{\alpha}}^{(n)}(s) \coloneqq \frac{1}{\lambda(\mathcal{T}_k^{(n)})} \int_{\mathcal{T}_k^{(n)}} \boldsymbol{\alpha}(\sigma)\,\mathrm{d}\sigma \text{ for } s \in \mathcal{T}_k^{(n)}.$$

As $\boldsymbol{\alpha}$ is a relaxed control, the $\widetilde{\boldsymbol{\alpha}}^{(n)}$ are relaxed controls as well. Furthermore, we have

$$\int_{\mathcal{T}_k^{(n)}} \widetilde{\boldsymbol{\alpha}}^{(n)}(s)\,\mathrm{d}s = \int_{\mathcal{T}_k^{(n)}} \boldsymbol{\alpha}(s)\,\mathrm{d}s$$

for all $n \in \mathbb{N}$ and all $k \in \{1, \dots, N^{(n)}\}$, in particular $d^{(n)}(\boldsymbol{\omega}^{(n)}, \boldsymbol{\alpha}) = d^{(n)}(\boldsymbol{\omega}^{(n)}, \widetilde{\boldsymbol{\alpha}}^{(n)})$.

We insert a zero into the duality pairing to rewrite it as

$$\left\langle \boldsymbol{\phi}_i^{(n)}, f \right\rangle_{L^\infty, L^1} = \int_{\Omega_T} (\widetilde{\boldsymbol{\alpha}}_i^{(n)} - \boldsymbol{\omega}_i^{(n)}) f + \int_{\Omega_T} (\boldsymbol{\alpha}_i - \widetilde{\boldsymbol{\alpha}}_i^{(n)}) f. \tag{7.3}$$

Combining (7.2) and Lemma 7.4, the first summand tends to zero. We have to show that the second summand tends to zero.

We employ the Lebesgue differentiation theorem, see [111, Chap. 3, Cor. 1.6 & 1.7], which can be applied due to the regular shrinkage assumption in Definition 7.6 for the admissible sequence of refined rounding grids and obtain

$$\widetilde{\boldsymbol{\alpha}}^{(n)}(s) \to \boldsymbol{\alpha}(s) \text{ for a.a. } s \in \Omega_T.$$

We employ Egorov's theorem, see [111, Chap. 1, Thm 4.4] to this pointwise a.e. convergence, which gives that for all $h > 0$, there exists $A_h \in \mathcal{B}(\Omega_T)$ such that

$$\widetilde{\boldsymbol{\alpha}}^{(n)}|_{A_h} \to \boldsymbol{\alpha}|_{A_h} \text{ in } L^\infty(\Omega_T) \tag{7.4}$$

and $\lambda(\Omega_T \backslash A_h) < \frac{h}{2}$. Furthermore, by the absolute continuity of the integral, see [111, Prop. 1.12 (ii)], for all $\delta > 0$, there exists $h > 0$ such that $\int_E |f| < \frac{\delta}{2}$ if $\lambda(E) < h$ for some $E \in \mathcal{B}(\Omega_T)$. Let $\varepsilon > 0$. Thus, let $h < \varepsilon$ be small enough such that $\lambda(E) < h$ implies $\int_E |f| < \frac{\varepsilon}{2}$ for the choice $E = \Omega_T \backslash A_h$. Consequently, we can estimate

$$\left| \int_{\Omega_T} (\boldsymbol{\alpha}_i - \widetilde{\boldsymbol{\alpha}}_i^{(n)}) f \right| - \left| \int_{A_h} (\boldsymbol{\alpha}_i - \widetilde{\boldsymbol{\alpha}}_i^{(n)}) f \right| \leq \left| \int_{\Omega_T \backslash A_h} (\boldsymbol{\alpha}_i - \widetilde{\boldsymbol{\alpha}}_i^{(n)}) f \right| \leq 2 \int_{\Omega_T \backslash A_h} |f| < \varepsilon.$$

The first inequality follows from the triangle inequality and the second from the uniform boundedness of $\boldsymbol{\alpha}_i$ and $\widetilde{\boldsymbol{\alpha}}_i^{(n)}$ by one and the third from the fact that $h < \varepsilon$ was chosen small enough. Let $h > 0$ be fixed. Then, the subtracted term tends to zero by virtue of (7.4). As $\varepsilon > 0$ was chosen arbitrarily, the second summand in (7.3) tends to zero as well. This closes the proof. □

As for Lemma 7.4, we note that the convergence holds regardless of permutations of the grid cells if (SUR) or (SUR-VC) is employed for rounding by the findings in Chapter 6. The algorithms also ensure the prerequisite that the $\boldsymbol{\omega}^{(n)}$ are a.e. constant on all grid cells. Thus, Lemma 7.8 yields the desired approximation $\boldsymbol{\omega}^{(n)} \rightharpoonup \boldsymbol{\alpha}$, which we provide below.

Theorem 7.9 ([79, Thm 4.7]). *Let* $\left(\left\{\mathcal{T}_1^{(n)}, \ldots, \mathcal{T}_{N^{(n)}}^{(n)}\right\}\right)_n$ *be an admissible sequence of refined rounding grids of a bounded domain* $\Omega_T \subset \mathbb{R}^d$ *with corresponding integrality gaps* $(d^{(n)})_n$*. Let* $\boldsymbol{\alpha}$ *be a relaxed control and* $(\boldsymbol{\omega}^{(n)})_n$ *be a sequence of binary controls such that* $d^{(n)}(\boldsymbol{\omega}^{(n)}, \boldsymbol{\alpha}) \to 0$*. Then,*

$$\boldsymbol{\phi}^{(n)} \rightharpoonup 0 \text{ in } L^p(\Omega_T, \mathbb{R}^M) \text{ for } 1 \leq p < \infty$$

and

$$\boldsymbol{\phi}^{(n)} \rightharpoonup^* 0 \text{ in } L^p(\Omega_T, \mathbb{R}^M) \text{ for } 1 < p \leq \infty.$$

Proof. We employ the identification $\mathbb{R}^M \cong (\mathbb{R}^M)^*$ and L^p space duality for vector-valued function spaces, see Theorem B.27, together with Lemma 7.8. □

Chapter 8

Topological characterization of the integrality gap

Although the following observations do not have any important consequences for this work and are mostly derived for the one-dimensional case, in which the integrality gap can be interpreted independent of the grid, they help to characterize the approximation achieved by rounding algorithms that satisfy (a) in (1.1).

Let $\boldsymbol{\alpha}$, $\boldsymbol{\omega} \in L^\infty((0,T))$ and recall $d_{\mathsf{1D}} : L^\infty((0,T)) \times L^\infty((0,T)) \to \mathbb{R}$

$$d_{\mathsf{1D}}(\boldsymbol{\omega}, \boldsymbol{\alpha}) = \sup_{t\in[0,T]} \left| \int_0^t \boldsymbol{\alpha} - \boldsymbol{\omega} \,\mathrm{d}\lambda \right|$$

for the one-dimensional case. The following proposition follows from Section 7.1.

Proposition 8.1 (Theorem 7.3)**.** *Let the sequence $(\boldsymbol{\omega}^{(n)})_n \subset L^\infty((0,T))$ be bounded, $\boldsymbol{\alpha} \in L^\infty((0,T))$ and $d_{\mathsf{1D}}(\boldsymbol{\omega}^{(n)}, \boldsymbol{\alpha}) \to 0$. Then, $\boldsymbol{\omega}^{(n)} \overset{*}{\rightharpoonup} \boldsymbol{\alpha}$.* ■

But interestingly, a more precise characterization exists. We consider the mapping $\nu : L^\infty((0,T)) \to [0, +\infty)$

$$\nu(\boldsymbol{\phi}) \coloneqq \sup_{t\in[0,T]} \left| \int_0^t \boldsymbol{\phi} \,\mathrm{d}\lambda \right|$$

and realize that it defines a norm on $L^\infty((0,T))$ as the following proposition shows.

Proposition 8.2. *ν is a norm on $L^\infty((0,T))$.*

Proof. The absolute homogeneity is obvious. The triangle inequality follows from linearity of the integral, the triangle inequality of $|\cdot|$ and the subadditivity of $\sup$. Let $\boldsymbol{\phi} \in L^\infty((0,T))$ satisfy $\nu(\boldsymbol{\phi}) = 0$. Then, for any $0 < a < b < T$, we obtain

$$\int_a^b \boldsymbol{\phi} \,\mathrm{d}\lambda = \int_0^b \boldsymbol{\phi} \,\mathrm{d}\lambda - \int_0^a \boldsymbol{\phi} \,\mathrm{d}\lambda = 0.$$

We continue with a contradiction argument. Suppose $\boldsymbol{\phi} > 0$ on $A \subset [a,b]$ with $\lambda(A) > 0$. Then, there exists a closed set $B \subset A$ with $\lambda(B) > 0$. Then, the set $(a,b)\backslash B$ is open

and hence a countable disjoint union of open intervals. But $\boldsymbol{\phi}$ vanishes on intervals. Thus, $\boldsymbol{\phi} = 0$ on $(a,b)\backslash B$. But this means that

$$0 = \int_a^b \boldsymbol{\phi}\,\mathrm{d}\lambda = \int_B \boldsymbol{\phi}\,\mathrm{d}\lambda + \int_{(a,b)\backslash B} \boldsymbol{\phi}\,\mathrm{d}\lambda = \int_B \boldsymbol{\phi}\,\mathrm{d}\lambda > 0,$$

which is a contradiction and closes the argument. □

We observe immediately that d_{1D} is the metric induced by ν. We continue to characterize convergence in $L^\infty((0,T))$ w.r.t. ν. We begin by showing that convergence in ν is weaker than sequential convergence in $\sigma(L^\infty, L^1)$.

Lemma 8.3. *Let $(\boldsymbol{\omega}^{(n)})_n \subset L^\infty((0,T))$ and $\boldsymbol{\alpha} \in L^\infty((0,T))$ satisfy $\boldsymbol{\omega}^{(n)} \rightharpoonup^* \boldsymbol{\alpha}$. Then,*

$$\nu(\boldsymbol{\alpha} - \boldsymbol{\omega}^{(n)}) \to 0.$$

Proof. For all $t \in (0,T)$, we have $\chi_{(0,t)} \in L^1((0,T))$ and deduce that $\left|\int_0^t \boldsymbol{\alpha} - \boldsymbol{\omega}^{(n)}\right| \to 0$ holds pointwise. Next, we check the prerequisites of the Arzelà-Ascoli theorem for the family of functions $F \coloneqq \{\boldsymbol{\Phi}_{\mathsf{1D}}^{(n)} : n \in \mathbb{N}\}$ with $\boldsymbol{\Phi}_{\mathsf{1D}}^{(n)}(t) \coloneqq \int_0^t \boldsymbol{\alpha} - \boldsymbol{\omega}^{(n)}$. As weak* convergent sequences are bounded w.r.t. $\|\cdot\|_{L^\infty}$, we obtain $\|\boldsymbol{\alpha} - \boldsymbol{\omega}^{(n)}\|_{L^\infty} \leq C$ for some C and all $n \in \mathbb{N}$ and consequently, uniform boundedness of the $\boldsymbol{\Phi}_{\mathsf{1D}}^{(n)}$. Furthermore, we consider $a, b \in (0,T)$ with $|a-b| \leq \delta$. Then,

$$\left|\boldsymbol{\Phi}_{\mathsf{1D}}^{(n)}(b) - \boldsymbol{\Phi}_{\mathsf{1D}}^{(n)}(a)\right| = \left|\int_a^b \boldsymbol{\alpha} - \boldsymbol{\omega}^{(n)}\,\mathrm{d}\lambda\right| \leq \delta C,$$

which establishes uniform equicontinuity of the $\boldsymbol{\Phi}_{\mathsf{1D}}^{(n)}$. Thus, $F \subset\subset C([0,T])$ and every subsequence of $(\boldsymbol{\Phi}_{\mathsf{1D}}^{(n)})_n$ has a uniformly convergent subsequence. But as all subsubsequences converge to zero pointwise, the limits coincide and are equal to 0. Thus, $\boldsymbol{\Phi}_{\mathsf{1D}}^{(n)} \to 0$ in $C([0,T])$. This yields the claim. □

Lemma 8.4. *Let $(\boldsymbol{\omega}^{(n)})_n \subset L^\infty((0,T))$ be bounded w.r.t. $\|\cdot\|_{L^\infty}$ and let $\boldsymbol{\alpha} \in L^\infty((0,T))$. Let $\nu(\boldsymbol{\omega}^{(n)} - \boldsymbol{\alpha}) \to 0$. Then, $\boldsymbol{\omega}^{(n)} \rightharpoonup^* \boldsymbol{\alpha}$.*

Proof. We give the following alternative topological proof (the claim is the same as the one of Proposition 8.1). The boundedness implies the existence of a weak* convergent subsequence with limit $\boldsymbol{\beta}$ and by Lemma 8.3, the subsequence ν-converges to $\boldsymbol{\beta}$. But the prerequisites tell us that it ν-converges to $\boldsymbol{\alpha}$ and by uniqueness of the limit in metric spaces, the limits coincide. Passing to subsubsequences, we obtain $\boldsymbol{\omega}^{(n)} \rightharpoonup^* \boldsymbol{\alpha}$. □

Boundedness in $\|\cdot\|_{L^\infty}$ is necessary for Lemma 8.4 as the following example shows.

Example 8.5. *Let $T = 1$ and consider the indicator function*

$$i_n(x) \coloneqq \begin{cases} 1 & \text{if } x \in [\frac{m-1}{n^2}, \frac{m}{n^2}) \text{ for } m = 2k \text{ and } 1 \leq k \leq \left\lfloor \frac{n^2}{2} \right\rfloor, \\ -1 & \text{else,} \end{cases}$$

which is 1 *on* $\lfloor n^2/2 \rfloor$ *and* -1 *on* $\lceil n^2/2 \rceil$ *pairwise disjoint intervals of measure* $1/n^2$ *that partition* $[0,T]$*. We consider the sequence of the functions defined by* $\boldsymbol{\phi}^{(n)}(x) \coloneqq i_n(x)n$*. Then,*

$$\sup_{t\in[0,T]} \left| \int_0^t \boldsymbol{\phi}^{(n)} \,\mathrm{d}\lambda \right| = n\frac{1}{n^2} = \frac{1}{n} \to 0.$$

But the sequence is $\|\cdot\|_{L^\infty}$*-unbounded and therefore, it cannot converge in* $\sigma(L^\infty, L^1)$*.*

Similar arguments to the ones above hold for all L^p spaces.

Lemma 8.6. $(L^p((0,T)),\nu)$ *is not complete for* $1 < p \leq \infty$*.*

Proof. We observe that $\nu(\boldsymbol{\phi}) \leq \|\boldsymbol{\phi}\|_{L^1}$ for all $\boldsymbol{\phi} \in L^1$. Let $(\boldsymbol{\omega}^{(n)})_n \subset L^p((0,T))$ with $(\boldsymbol{\omega}^{(n)})_n \to \boldsymbol{\alpha}$ in L^1. Such sequences exist by denseness of the embeddings $L^p((0,T)) \hookrightarrow L^q((0,T))$ for $p > q$. Then, $\nu(\boldsymbol{\omega}^{(n)} - \boldsymbol{\alpha}) \to 0$ and thus, $\boldsymbol{\omega}^{(n)}$ is Cauchy w.r.t. ν with a limit not necessarily in $L^p((0,T))$. □

For $\boldsymbol{\phi} \in L^1((0,T))$, we define the measure induced by $\boldsymbol{\phi}$ as $\mu|_\phi(A) \coloneqq \int_A \boldsymbol{\phi} \,\mathrm{d}\lambda$, i.e. $\boldsymbol{\phi}$ is the Radon-Nikodym derivative of $\mu|_\phi$ w.r.t. λ. We consider the space $\mathfrak{M}([0,T],\mathcal{B})$ of regular, bounded, countably additive, signed measures μ equipped with the total variation norm $\|\mu\|_{\mathfrak{M}} = |\mu|([0,T])$. We define ν for measures as follows

$$\nu(\mu) \coloneqq \sup_{t\in[0,T]} |\mu([0,t])| \,.$$

Proposition 8.7. ν *is a norm on* $\mathfrak{M}$*. For* $\boldsymbol{\phi} \in L^1((0,T))$*, we have* $\nu(\mu|_\phi) = \nu(\boldsymbol{\phi})$*.*

Proof. The only part of the proof that is not immediate is the positive definiteness of ν. First, we consider the following generator of $\mathcal{B}$.

$$\mathcal{E} \coloneqq \{[0,t] : t \in [0,T]\},$$

which is in particular closed under finite intersections. Let $\nu(\mu) = 0$. Then, $\mu(A) = 0$ for all $A \in \mathcal{E}$. We consider the set

$$\mathcal{D} \coloneqq \{A \in \mathcal{B} : \mu(A) = 0\}.$$

From $\nu(\mu) = 0$, we obtain $\mu([0,T]) = 0$ and verify straightforwardly that $\mathcal{D}$ is a Dynkin system. As $\mathcal{E}$ is closed under finite intersections, we may employ Dynkin's π-λ-theorem and obtain that the Dynkin system generated by $\mathcal{E}$ is identical to $\mathcal{B}$. As $\mathcal{E} \subset \mathcal{D}$, we obtain $\mathcal{B} \subset \mathcal{D}$ and thus $\mathcal{B} = \mathcal{D}$. Consequently, $\mu(A) = 0$ for all $A \in \mathcal{B}$, i.e. $\mu = 0$, which proves the claim. □

We briefly investigate the relationship of ν to the total variation norm.

Lemma 8.8. *For all* $\boldsymbol{\phi} \in L^p((0,T))$*, we have*

$$\nu(\boldsymbol{\phi}) \leq \|\mu|_\phi\|_{\mathfrak{M}}.$$

and for all $\mu \in \mathfrak{M}([0,T],\mathcal{B})$*, we have*

$$\nu(\mu) \leq \|\mu\|_{\mathfrak{M}}.$$

Proof. For all $t \in \mathbb{R}$, the decomposition of $[0,T]$ into $[0,t]$, $(t,T]$ satisfies

$$\nu(\boldsymbol{\phi}) = \sup_t |\mu|_{\boldsymbol{\phi}}([0,t])| \leq \sup_t |\mu|_{\boldsymbol{\phi}}([0,t])| + |\mu|_{\boldsymbol{\phi}}((t,T])| \leq \|\mu|_{\boldsymbol{\phi}}\|_{\mathfrak{M}}.$$

The second claim follows analogously. □

Lemma 8.9. *There is no $c > 0$ such that $c\|\mu|_{\boldsymbol{\phi}}\|_{\mathfrak{M}} \leq \nu(\boldsymbol{\phi})$ for all $\boldsymbol{\phi} \in L^p((0,T))$.*

Proof. The claim follows from Example 8.5 and the fact that the induced measures are absolutely continuous w.r.t. λ and their total variation norms coincide with the L^1-norms for sequences of L^1-functions, i.e.

$$\|\mu|_{\boldsymbol{\phi}^{(n)}}\|_{\mathfrak{M}} = \|\boldsymbol{\phi}^{(n)}\|_{L^1} = Tn \to \infty$$

but $\nu(\boldsymbol{\phi}^{(n)}) \to 0$. □

Lemma 8.10. *$\overline{L^1((0,T))}^{\nu}$ is not continuously embedded in $\mathfrak{M}([0,T],\mathcal{B}) \cong C([0,T])^*$. In particular, $L^1((0,T))$ and $\mathfrak{M}([0,T],\mathcal{B})$ are not complete w.r.t. ν.*

Proof. Consider $\boldsymbol{\phi}^{(n)} := \frac{3^n}{2^n} i_n(x)$ with

$$i_n(x) := \begin{cases} 1 & \text{if } x \in [\frac{m-1}{2^{n+1}}, \frac{m}{2^{n+1}}) \text{ for } m = 2k \text{ and } 1 \leq k \leq 2^n, \\ -1 & \text{else.} \end{cases}$$

This gives $\nu(\boldsymbol{\phi}^{(n)}) \leq \frac{3^n}{4^n 2} \to 0$. We consider the Weierstraß function

$$w : x \mapsto \sum_{k=1}^{\infty} \frac{2^k}{3^k} \sin(2^{k+1}\pi x),$$

which is in $C([0,T]) \backslash C^1([0,T])$. We observe

$$\int_0^1 \boldsymbol{\phi}^{(n)} w \,\mathrm{d}\lambda = \int_0^1 \frac{3^n}{2^n}\frac{2^n}{3^n} |\sin(2^{n+1}\pi x)| \,\mathrm{d}\lambda \geq \frac{1}{2}$$

as i_n masks all but one summand of the Weierstraß series for every n, see [80, Example 2.2]. Thus, $(\boldsymbol{\phi}^{(n)})_n$ is a sequence of linear functionals on $C([0,T])$ and Cauchy w.r.t. ν, but it does not converge to a bounded linear functional on $C([0,T])$. □

Lemma 8.11. *$\overline{L^1((0,T))}^{\nu} \hookrightarrow^c C^1([0,T])^*$ and $\overline{L^1((0,T))}^{\nu} \hookrightarrow^c W^{1,\infty}((0,T))^*$.*

Proof. Let $f \in W^{1,\infty}((0,T))$ and $(\boldsymbol{\phi}^{(n)})_n \subset L^1$ be Cauchy w.r.t. ν. Then, integration by parts and $W^{1,\infty}((0,T)) \hookrightarrow^c C([0,T])$ give

$$\begin{aligned} \left|\langle \boldsymbol{\phi}^{(n)}, f\rangle_{(W^{1,\infty})^*, W^{1,\infty}}\right| &= \left|\int_0^T \boldsymbol{\phi}^{(n)}(s) \,\mathrm{d}s\, f(T) - \int_0^T \int_0^t \boldsymbol{\phi}^{(n)}(s) \,\mathrm{d}s\, f(t) \,\mathrm{d}t\right| \\ &\leq C\nu(\boldsymbol{\phi}^{(n)}) \left(\|f\|_{L^\infty} + \|f'\|_{L^\infty}\right) \end{aligned}$$

for $C = \max\{1, T\}$. We obtain

$$\|\boldsymbol{\phi}^{(n)}\|_{(W^{1,\infty})^*} = \sup\{\langle \boldsymbol{\phi}^{(n)}, f\rangle_{(W^{1,\infty})^*,W^{1,\infty}} : \|f\|_{W^{1,\infty}} \leq 1\} \leq C\nu(\boldsymbol{\phi}^{(n)}). \tag{8.1}$$

Thus, for any $\boldsymbol{\phi} \in \overline{L^1((0,T))}^{\nu}$ and $f \in W^{1,\infty}$, we may define

$$\langle \boldsymbol{\phi}, f\rangle_{(W^{1,\infty})^*,W^{1,\infty}} \coloneqq \lim \int_0^T \boldsymbol{\phi}^{(n)} \,\mathrm{d}s f(T) - \int_0^T \int_0^t \boldsymbol{\phi}^{(n)} \,\mathrm{d}s f' \,\mathrm{d}t$$

for $\boldsymbol{\phi}^{(n)} \to \boldsymbol{\phi}$ in ν with $(\boldsymbol{\phi}^{(n)})_n \subset L^1((0,T))$. For well-definedness, we verify that the limit does not depend on the choice of the approximating sequence. To this end, let $\boldsymbol{\psi}^{(n)} \to \boldsymbol{\phi}$ in ν. We obtain

$$\begin{aligned}&\left|\int_0^T \boldsymbol{\phi}^{(n)} \,\mathrm{d}s f(T) - \int_0^T \int_0^t \boldsymbol{\phi}^{(n)} \,\mathrm{d}s f' \,\mathrm{d}t - \int_0^T \boldsymbol{\psi}^{(n)} \,\mathrm{d}s f(T) - \int_0^T \int_0^t \boldsymbol{\psi}^{(n)} \,\mathrm{d}s f' \,\mathrm{d}t\right| \\ &\quad \leq C\nu(\boldsymbol{\phi}^{(n)} - \boldsymbol{\psi}^{(n)})\|f\|_{W^{1,\infty}} \to 0.\end{aligned}$$

as the limit of the sum of the two sequences is the sum of the limits by virtue of the triangle inequality. □

Lemma 8.12. $\overline{\mathfrak{M}([0,T],\mathbb{R})}^{\nu} \hookrightarrow^c W^{1,\infty}((0,T))^*$.

Proof. Let $f \in W^{1,\infty}$. We approximate f with piecewise constant functions. Let $0 = t_0 < \ldots < t_n = T$ partition $[0,T]$ into n intervals of length T/n. For $n \in \mathbb{N}$ and $i \in \{1, \ldots, n\}$, we define

$$f_i \coloneqq \frac{n}{T} \int_{t_{i-1}}^{t_i} f(t) \,\mathrm{d}t$$

and $f^{(n)} \coloneqq \sum_{i=1}^n f_i \chi_i$ with $\chi_i \coloneqq \chi_{(t_{i-1},t_i]}$ for $i > 1$ and $\chi_1 \coloneqq \chi_{[t_0,t_1]}$. We observe

$$\min_{t\in[t_{i-1},t_i]} f(t) \leq f_i \leq \max_{t\in[t_{i-1},t_i]} f(t)$$

and the intermediate value theorem gives $f_i = f(s_i)$ for some $s_i \in [t_{i-1}, t_i]$. Combining this with the Lipschitz continuity of f, we obtain

$$\left|\int_0^T |f(t) - \sum_{i=1}^n f_i \chi_i(t)| \,\mathrm{d}\mu(t)\right| = \left|\sum_{i=1}^n \int_{t_{i-1}}^{t_i} |f(t) - f(s_i)| \,\mathrm{d}\mu(t)\right| \leq \|f'\|_{L^\infty}\|\mu\|_{\mathfrak{M}}\frac{T}{n}.$$

for all $\mu \in \mathfrak{M}$. Let $(\mu^{(m)})_m \subset \mathfrak{M}$ be Cauchy w.r.t. ν. Then, we can reformulate

$$\begin{aligned}\left|\left\langle \mu^{(m)}, f\right\rangle_{\mathfrak{M},C}\right| &= \lim_n \left|\int_0^T \sum_{i=1}^n \chi_i(t) f_i \,\mathrm{d}\mu^{(m)}(t)\right| \\ &= \lim_n \left|f_1\mu^{(m)}([0,t_1]) + \sum_{i=2}^n f_i\mu^{(m)}((t_{i-1},t_i])\right| \\ &= \lim_n \left|\sum_{i=1}^{n-1}(f_i - f_{i+1})\mu^{(m)}([0,t_i]) + f_n\mu^{(m)}([0,T])\right|.\end{aligned}$$

As above, we obtain $|f_i - f_{i+1}| \leq \|f'\|_{L^\infty}(2T/n)$ and may estimate

$$\begin{aligned}\left|\left\langle \mu^{(m)}, f\right\rangle_{\mathfrak{M},C}\right| &\leq \lim_n \sup_{t\in[0,T]} |\mu^{(m)}([0,t])|2T\|f'\|_{L^\infty}\frac{n-1}{n} + \|f\|_{L^\infty}|\mu^{(m)}([0,T])| \\ &\leq \nu(\mu^{(m)})2\max\{1,T\}(\|f\|_{L^\infty} + \|f'\|_{L^\infty}),\end{aligned}$$

which proves the claim. □

Lemma 8.13. *There is no $c > 0$ such that $c\nu(\mu) \leq \sup\{\langle \mu, f\rangle_{\mathfrak{M},C} : \|f\|_{W^{1,\infty}} \leq 1\}$ for all $\mu \in \mathfrak{M}([0,T],\mathcal{B})$.*

Proof. We consider the sequence $(\delta_s - \delta_{s+\frac{1}{n}})_n \subset \mathfrak{M}$ for some $s \in (0,T)$. Then, we obtain

$$|\langle \delta_s - \delta_{s+\frac{1}{n}}, f\rangle_{\mathfrak{M},C}| = \left|f(s) - f\left(s+\frac{1}{n}\right)\right| \leq \|f\|_{W^{1,\infty}}\frac{1}{n} \to 0$$

for $f \in W^{1,\infty}$. However,

$$\nu\left(\delta_s - \delta_{s+\frac{1}{n}}\right) \geq \left|\delta_s\left(\left[0, s+\frac{1}{2n}\right]\right) - \delta_{s+\frac{1}{n}}\left(\left[0, s+\frac{1}{2n}\right]\right)\right| = 1.$$

□

We summarize our findings in the following theorem.[6]

Theorem 8.14. *Let $\phi^{(n)}$ be a sequence of measurable functions. Then,*

1. $\phi^{(n)} \to_{\|\cdot\|_p} \phi \quad\Rightarrow\quad \phi^{(n)} \rightharpoonup_p \phi \quad\Rightarrow\quad \phi^{(n)} \to_\nu \phi \quad$ *for* $1 \leq p < \infty$,
2. $\phi^{(n)} \to_{\|\cdot\|_p} \phi \quad\Rightarrow\quad \phi^{(n)} \rightharpoonup^*_p \phi \quad\Rightarrow\quad \phi^{(n)} \to_\nu \phi \quad$ *for* $1 < p \leq \infty$.

Let $(\mu^{(n)})_n \subset \mathfrak{M}([0,T],\mathcal{B})$. Then,

3. $\mu^{(n)} \to_{\|\cdot\|_{\mathfrak{M}}} \mu \quad\Rightarrow\quad \mu^{(n)} \to_\nu \mu \quad\Rightarrow\quad \begin{array}{l}\mu^{(n)} \to_{(C^1)^*} \mu, \\ \mu^{(n)} \to_{(W^{1,\infty})^*} \mu.\end{array}$

All reverse implications are false.

Finally, we consider the multi-dimensional case.

Lemma 8.15. *Let a rounding grid for a bounded domain $\Omega_T \subset \mathbb{R}^d$ be given and $p \in [1,\infty]$. Then, the mapping $d : L^p(\Omega_T,\mathbb{R}^M) \times L^p(\Omega_T,\mathbb{R}^M) \to \mathbb{R}$ defined in Definition 6.1 is a pseudometric.*

Proof. The symmetry follows from the symmetry of $\|\cdot\|_\infty$ on $\mathbb{R}^M$. The triangle inequality follows from the linearity of the integral, the triangle inequality of $\|\cdot\|_\infty$ on $\mathbb{R}^M$ and the subadditivity of $\operatorname{ess\,sup}$. □

[6] We acknowledge that several insights and ideas for the proofs in this section were developed in discussions with Dirk Lorenz, TU Braunschweig.

Remark 8.16. *As introduced in Definition 6.1, the codomain space $\mathbb{R}^M$ is equipped with $\|\cdot\|_\infty$ in this work to analyze the integrality gap and establishing its properties. As norms are equivalent on $\mathbb{R}^M$, other norms on the codomain do not change* (1.1).

Proposition 8.17. *Consider an admissible sequence of rounding grids, see Definition 7.6. Then, the induced family of functions*

$$\nu^{(n)}(\mu) \coloneqq \max_{k\in\{1,\dots,N^{(n)}\}} \left|\mu\left(\bigcup_{\ell=1}^{k} \mathcal{T}_\ell^{(n)}\right)\right|$$

on signed Borel measures $\mu : \mathcal{B}(\bar{\Omega}) \to \mathbb{R}$ constitutes a family of seminorms. We define the set $\mathcal{E} \subset 2^{\bar{\Omega}}$ as

$$\mathcal{E} \coloneqq \bigcup_{n=1}^{\infty} \left\{\mathcal{T}_1^{(n)}, \dots, \mathcal{T}_{N^{(n)}}^{(n)}\right\}.$$

Let $\mathcal{E}$ be closed under finite intersections and be a generator of $\mathcal{B}(\bar{\Omega})$. Then, the locally convex vector space of the signed Borel measures with the topology determined by the family of seminorms $(\nu^{(n)})_n$ is Hausdorff.

Proof. The seminorm properties follow similar to Lemma 8.15. Thus, the vector space of signed Borel measures with the topology determined by the family of seminorms $(\nu^{(n)})_n$ is locally convex. To prove the Hausdorff property, we verify $(\forall n : \nu^{(n)}(\mu) = 0) \Rightarrow \mu = 0$. Let $\nu^{(n)}(\mu) = 0$ for all $n \in \mathbb{N}$. Then, Definition 7.6 asserts $\mu(\bar{\Omega}) = 0$. We consider the set

$$\mathcal{D} \coloneqq \{A \in \mathcal{B} : \mu(A) = 0\}.$$

Using $\mu(\bar{\Omega}) = 0$, it follows straightforwardly that $\mathcal{D}$ is a Dynkin system. We have $\mathcal{E} \subset \mathcal{D}$ because $\nu^{(n)}(\mu) = 0$ for all $n \in \mathbb{N}$. Indeed, we have

$$\mu(\mathcal{T}_k^{(n)}) = \mu\left(\bigcup_{\ell=1}^{k} \mathcal{T}_\ell^{(n)}\right) - \mu\left(\bigcup_{\ell=1}^{k-1} \mathcal{T}_\ell^{(n)}\right) = 0$$

for all $k \in N^{(n)}$ for all $n \in \mathbb{N}$. Consequently, $\mathcal{E} \subset \mathcal{D}$ and thus, the Dynkin system generated by $\mathcal{E}$ is a subset of $\mathcal{D}$. As $\mathcal{E}$ is closed under finite intersections, we may employ Dynkin's π-λ-theorem, see e.g. [14, Thm 1.9.3 (ii)], to deduce $\mathcal{D} = \mathcal{B}$. Consequently, $\mu(A) = 0$ for all $A \in \mathcal{B}$, i.e. $\mu = 0$, which proves the claim. □

Chapter 9

State vector approximation

This chapter serves to achieve contribution (c) in (1.1), i.e. we prove

$$\boldsymbol{\omega}^{(n)} \rightharpoonup \boldsymbol{\alpha} \quad \underset{(c)}{\Rightarrow} \quad y(\boldsymbol{\omega}^{(n)}) \to y(\boldsymbol{\alpha})$$

for a binary control sequence $(\boldsymbol{\omega}^{(n)})_n$, a relaxed control $\boldsymbol{\alpha}$ and the corresponding solutions of the state equations $(y(\boldsymbol{\omega}^{(n)}))_n$ and $y(\boldsymbol{\alpha})$. The first three sections prove (c) for the three classes of (MIPDECO) introduced in Chapter 3 in Theorems 9.3, 9.5 and 9.9 while the fourth section proves (c) for a class of convolution operators in Theorem 9.11. This illustrates that the findings are not restricted to differential equations and generalize to control problems that feature completely continuous control-to-state operators.

9.1 Distributed integer controls in time

This section considers the state vector approximation for the evolution problems constraining (MIPEVO-T), i.e.

$$\begin{aligned} \partial_t y + Ay &= -f(y, u, \boldsymbol{v}) \\ y(0) &= y_0 \end{aligned} \qquad \text{(3.2 revisited)}$$

under Assumption 3.1, in which the discrete control $\boldsymbol{v}$ is only time-dependent. We follow the arguments by the author, which have been developed in [80, Sect. 2.2]. The arguments directly connect to the arguments in Section 7.1. Before the state vector convergence can be shown, we need two results that are proven in the following two lemmata. We begin with a relationship of pointwise and uniform convergence.

Lemma 9.1 ([80, Lem. 2.5])**.** *Let Y be a Banach space, $(S(t))_{t\geq 0}$ be a strongly continuous semigroup on Y, $f \in L^1((0,T),Y)$. Then,*

$$\sup_{t\in[0,T]} \int_0^t \|(S(t+h-s) - S(t-s))f(s)\|_Y \,\mathrm{d}s \to 0$$

for $h \downarrow 0$.

Proof. First, we observe that there exists an exponential function that dominates the mapping $t \mapsto \|S(t)\|_{op}$, see Proposition B.8. As we are restricted to the compact interval $[0,T]$, this allows us to define $C \coloneqq \sup_{t\in[0,T]} \|S(t)\|_{op}$. We estimate

$$\begin{aligned}
&\sup_{t\in[0,T]} \int_0^t \|(S(t+h-s) - S(t-s))f(s)\|_Y \,\mathrm{d}s \\
&\quad \leq \sup_{t\in[0,T]} \int_0^t \|S(t-s)\|_{op} \|(S(h) - I)f(s)\|_Y \,\mathrm{d}s \\
&\quad \leq C \int_0^T \|(S(h) - I)f(s)\|_Y \,\mathrm{d}s,
\end{aligned}$$

where the first inequality follows from the submultiplicativity of the operator norm and the semigroup property of $(S(t))_{t\geq 0}$. The claim follows from Lebesgue's dominated convergence theorem. □

The second preparatory lemma reveals that the elements of certain sequences in $C([0,T],Y)$ constitute relatively compact sets.

Lemma 9.2 ([80, Lem. 2.6])**.** *Let Y be a Banach space, $(S(t))_{t\geq 0}$ be a strongly continuous semigroup on Y, $f \in L^1((0,T),Y)$, $(\phi^{(n)})_n \subset L^\infty((0,T),\mathbb{R})$ with $\phi^{(n)}(t) \in [-1,1]$ for a.a. $t \in [0,T]$ such that*

$$\nu^{(n)}(t) \to 0 \text{ for all } t \in [0,T]$$

with the definition

$$\nu^{(n)}(t) \coloneqq \int_0^t \phi^{(n)}(s) S(t-s) f(s) \,\mathrm{d}s.$$

Then, the set $\left\{\nu^{(n)} : n \in \mathbb{N}\right\}$ is relatively compact in $L^p((0,T),Y)$ for $p \in [1,\infty)$ and $C([0,T],Y)$ in the norm topology.

Proof. Again by virtue of Proposition B.8, we define $C \coloneqq \sup_{t\in[0,T]} \|S(t)\|_{op} < \infty$. The absolute continuity of the Bochner integral, see Proposition B.22, implies $(\nu^{(n)})_n \subset C([0,T],Y)$. The boundedness of the sequence $(\phi^{(n)})_n$ and the boundedness of $(S(t))_{t\geq 0}$ in $[0,T]$ give boundedness of $(\nu^{(n)})_n$. The claim follows if [108, Thm 1] by Simon can be applied, which can be interpreted as an application and extension of the Arzelà-Ascoli theorem and is an ingredient of the Aubin-Lions-Simon lemma. To this end, we have to verify the conditions

$$B_{t_1,t_2} \coloneqq \left\{ \int_{t_1}^{t_2} \nu^{(n)}(t) \,\mathrm{d}t : n \in \mathbb{N} \right\} \subset\subset Y \text{ for all } 0 < t_1 < t_2 < T \tag{9.1}$$

and

$$\sup_{n\in\mathbb{N}} \|\nu^{(n)}(\cdot + h) - \nu^{(n)}(\cdot)\|_{L^p((0,T-h),Y)} \to 0 \text{ for } h \downarrow 0 \tag{9.2}$$

to show convergence in $L^p((0,T),Y)$ for $p<\infty$, and in $C([0,T],Y)$ in the case $p=\infty$.

We begin with condition (9.1), which is easier to verify. We observe that $\nu^{(n)}(t)\to 0$ pointwise and the uniform boundedness of sequence $(\nu^{(n)})_n$. Thus, Lebesgue's dominated convergence theorem gives

$$\left\|\int_{t_1}^{t_2}\nu^{(n)}(t)\,\mathrm{d}t\right\|_Y\to 0$$

for all $0<t_1<t_2<T$. Consequently, the set B_{t_1,t_2} consists of the elements of a Cauchy sequence and therefore constitutes a relatively compact subset of Y.

To verify (9.2), we first estimate

$$\begin{aligned}
&\|\nu^{(n)}(t+h)-\nu^{(n)}(t)\|_Y\\
&\quad=\left\|\int_0^{t+h}S(t+h-s)f(s)\boldsymbol{\phi}^{(n)}(s)\,\mathrm{d}s-\int_0^t S(t-s)f(s)\boldsymbol{\phi}^{(n)}(s)\,\mathrm{d}s\right\|_Y\\
&\quad\le\left\|\int_t^{t+h}S(t+h-s)f(s)\boldsymbol{\phi}^{(n)}(s)\,\mathrm{d}s\right\|_Y\\
&\qquad+\left\|\int_0^t(S(t+h-s)-S(t-s))f(s)\boldsymbol{\phi}^{(n)}(s)\,\mathrm{d}s\right\|_Y
\end{aligned}$$

and consider the emerged summands separately. The first summand can be estimated from above by means of the triangle inequality and the bounds on $(\boldsymbol{\phi}^{(n)})_n$ and $(S(t))_{t\ge 0}$ in $[0,T]$, which gives

$$\left\|\int_t^{t+h}S(t+h-s)f(s)\boldsymbol{\phi}^{(n)}(s)\,\mathrm{d}s\right\|_Y\le C\int_0^T\|f(s)\|_Y\chi_{[t,t+h]}(s)\,\mathrm{d}s.$$

Immediately, convergence for $h\downarrow 0$ follows from Lebesgue's dominated convergence theorem independent of n. We continue with the second summand and estimate

$$\begin{aligned}
&\left\|\int_0^t(S(t+h-s)-S(t-s))f(s)\boldsymbol{\phi}^{(n)}(s)\,\mathrm{d}s\right\|_Y\\
&\qquad\le\int_0^t\|(S(t+h-s)-S(t-s))f(s)\|_Y\,|\boldsymbol{\phi}^{(n)}(s)|\,\mathrm{d}s\\
&\qquad\le\int_0^t\|(S(t+h-s)-S(t-s))f(s)\|_Y\,\mathrm{d}s.
\end{aligned}$$

Submultiplicativity of the operator norm and the existence of C allow the application of Lebesgue's dominated convergence theorem, which yields

$$\int_0^t\|(S(t+h-s)-S(t-s))f(s)\|_Y\,\mathrm{d}s\to 0\text{ for }h\downarrow 0\tag{9.3}$$

for all $t\in[0,T-h]$. This term is uniformly bounded for $h\in[0,T]$ and another application of Lebesgue's dominated convergence theorem gives

$$\int_0^{T-h}\left\|\int_0^t(S(t+h-s)-S(t-s))f(s)\,\mathrm{d}s\right\|_Y^p\,\mathrm{d}t\to 0\text{ for }h\downarrow 0$$

for all $p \in [1, \infty)$. The last case $p = \infty$, i.e. convergence in $C([0,T], Y)$, follows from the application of Lemma 9.1 to (9.3). Thus, (9.2) holds true for $\{\nu^{(n)} : n \in \mathbb{N}\}$. Consequently, [108, Thm 1] implies $\{\nu^{(n)} : n \in \mathbb{N}\}$ is relatively compact in the norm topology of $L^p((0,T), Y)$ for $p \in [1, \infty)$ and of $C([0,T], Y)$ in the case $p = \infty$. □

Next, we use Lemmas 7.1 and 9.2 to prove the state vector approximation for (3.2) and in turn (MIPEVO-T). We improve the work by Hante and Sager [50, 51] in the sense that we are able to drop assumptions on differentiability and uniformly bounded derivatives on the involved quantities from [51, Thm 1] and [50, Hyp. 3] to prove the convergence. The improvement comes at the cost of not being able to prove a priori estimates on the state vector approximation, which is done in [50, 51]. This is due to the fact that in the absence of the differentiability assumptions in [50, 51], we cannot employ integration by parts to obtain an estimate of the form $a(t) \leq C d_{1\mathrm{D}}(\boldsymbol{\omega}^{(n)}, \boldsymbol{\alpha})$ for some $C > 0$ on the term $a(t)$ in Theorem B.31 (Grönwall's inequality). Indeed, we only achieve $a(t) \to 0$ uniformly, but without any rate.

Theorem 9.3 ([80, Thm 2.7]). *Let $\Omega_T = (0,T)$ and let Assumption 3.1 hold. Let $\boldsymbol{\alpha}$ be a relaxed control, $(\boldsymbol{\omega}^{(n)})_n$ be a sequence of binary controls such that*

$$\sup_{t \in [0,T]} \left\| \int_0^t \boldsymbol{\phi}_i^{(n)}(s) f(s) \, \mathrm{d}s \right\|_Y \to 0$$

for $\boldsymbol{\phi}^{(n)} := \boldsymbol{\alpha} - \boldsymbol{\omega}^{(n)}$ and all $f \in L^1((0,T), Y)$ and $i \in \{1, \dots, M\}$. Let $y_0 \in Y$ and $u \in \mathcal{U}$. Let y be the unique mild solution of (3.1) for $\boldsymbol{\alpha}$ and let $y^{(n)}$ be the unique mild solutions of (3.3) for $\boldsymbol{\omega}^{(n)}$ and $n \in \mathbb{N}$. Then, there exists $C_r > 0$ such that for all $\varepsilon > 0$, there exist $n_0 \in \mathbb{N}$ such that the bound

$$\left\| y(t) - y^{(n)}(t) \right\|_Y \leq \varepsilon \exp(C_r t)$$

holds for all $n \geq n_0$.

Proof. We define $f_i := f(y, u, v_i)$, which is in $L^1((0,T), Y)$ by Assumption 3.1. Let $t \in [0,T]$. As mild solutions are continuous functions, we can evaluate them pointwise and use the variation of constants formulas (VOC) to compute

$$\|y(t) - y^{(n)}(t)\|_Y = \left\| \sum_{i=1}^M \int_0^t S(t-s)(\boldsymbol{\alpha}_i(s) f_i(s) - \boldsymbol{\omega}_i^{(n)}(s) f(y^{(n)}(s), u(s), v_i)) \, \mathrm{d}s \right\|_Y .$$

Let L denote the Lipschitz constant of f in the first argument. As in [100], we insert a

zero and estimate

$$\begin{aligned}
&\|y(t) - y^{(n)}(t)\|_Y \\
&\quad \leq \left\| \sum_{i=1}^{M} \int_0^t S(t-s)(\boldsymbol{\alpha}_i(s) f_i(s) - \boldsymbol{\omega}_i^{(n)}(s) f_i(s)) \,\mathrm{d}s \right\|_Y \\
&\qquad + \left\| \sum_{i=1}^{M} \int_0^t \boldsymbol{\omega}_i^{(n)}(s) S(t-s)(f_i(s) - f(y^{(n)}(s), u(s), v_i)) \,\mathrm{d}s \right\|_Y \\
&\quad \leq \sum_{i=1}^{M} \left\| \int_0^t \boldsymbol{\phi}_i^{(n)}(s) S(t-s) f_i(s) \,\mathrm{d}s \right\|_Y \\
&\qquad + ML \sup_{t\in[0,T]} \|S(t)\|_{op} \int_0^t \|y(s) - y^{(n)}(s)\|_Y \,\mathrm{d}s
\end{aligned}$$

using the bound $\|\boldsymbol{\omega}_i^{(n)}\|_{L^\infty((0,T))} \leq 1$. Again by virtue of Proposition B.8, $\|S(t)\|_{op}$ is bounded in $[0,T]$ and we define $C_r \coloneqq ML \sup_{t\in[0,T]} \|S(t)\|_{op}$.

To finish the proof, we want to apply Grönwall's inequality, see Theorem B.31. To this end, let $i \in \{1, \dots, M\}$ be fixed and consider the sequence

$$\nu_i^{(n)}(t) \coloneqq \int_0^t \boldsymbol{\phi}_i^{(n)}(s) S(t-s) f_i(s) \,\mathrm{d}s$$

for which we need to show

$$\sup_{t\in[0,T]} \|\nu_i^{(n)}(t)\|_Y \to 0.$$

The mapping $s \mapsto S(t-s) f_i(s)$ is in $L^1((0,t), Y)$, see Proposition B.30, which allows us to apply the prerequisites and obtain $\nu_i^{(n)}(t) \to 0$ for all $t \in [0,T]$. Furthermore, we have $(\boldsymbol{\phi}_i^{(n)})_n \subset L^\infty((0,T), \mathbb{R})$ with $\|\boldsymbol{\phi}_i^{(n)}\|_{L^\infty((0,T))} \leq 1$, which implies $(\nu_i^{(n)})_n \subset C([0,T], Y)$ and

$$\sup_{n\in\mathbb{N}} \sup_{t\in[0,T]} \|\nu_i^{(n)}(t)\|_Y < \infty,$$

where the continuity follows from the absolute continuity of the Bochner integral. Hence, $(\nu_i^{(n)})_n$ is a bounded sequence in $C([0,T], Y)$ that converges to 0 pointwise. By means of Dinculeanu and Singer's extension of the Riesz–Markov–Kakutani theorem, see Proposition B.19, elements ψ of the topological dual of $C([0,T], Y)$ can be identified with regular Borel measures with finite variation $\mu : \mathcal{B} \to Y^*$, which gives

$$\psi(\nu_i^{(n)}) \coloneqq \int_0^T \nu_i^{(n)} \,\mathrm{d}\mu.$$

Combining this, we may apply Lebesgue's dominated convergence theorem, see Proposition B.17, which yields

$$\nu_i^{(n)} \rightharpoonup 0.$$

Convergence in norm follows if $\{\nu_i^{(n)} : n \in \mathbb{N}\}$ is relatively compact w.r.t. the $\|\cdot\|_{C([0,T],Y)}$-topology because weakly convergent sequences contained in norm-compact

sets converge to the same limit in norm and sequential compactness is equivalent to compactness on metric spaces. Fortunately, we have already proven Lemma 9.2, from which we infer the desired compactness and combining this with $\nu_i^{(n)} \rightharpoonup 0$ gives

$$\lim_{n\to\infty} \|\nu_i^{(n)}\|_{C([0,T],Y)} = 0.$$

Finally, an $\frac{\varepsilon}{M}$-argument verifies that for all $\varepsilon > 0$ there exists $n_0 \in \mathbb{N}$ such that

$$\sup_{t\in[0,T]} \sum_{i=1}^{M} \left\| \int_0^t \boldsymbol{\phi}_i^{(n)}(s) S(t-s) f_i(s)\,\mathrm{d}s \right\|_Y < \varepsilon$$

for all $n \geq n_0$ and we finish the proof by virtue of Grönwall's inequality. □

Remark 9.4. *In Theorem 9.3, we show that $d_{1\mathrm{D}}(\boldsymbol{\omega}^{(n)}, \boldsymbol{\alpha}) \to 0$ implies convergence of the state vector in norm for a class of semilinear time-dependent PDEs without the differentiability assumptions imposed in* [51]. *As mentioned in Section 2.2, related ideas have been pursued from the 1960s onwards. Indeed, Warga's article* [119] *from 1975 pursues a similar goal in an ordinary differential inclusion setting. In* [119]*, a minimizer of a relaxed problem is approximated with so-called* admissible original solutions *and Pontryagin's maximum principle is transferred to the limiting minimizer without differentiability assumptions on the involved quantities, but under certain Lipschitz continuity assumptions. Similar to our proof, Warga employs smooth approximation and the identification of the topological dual of the continuous functions on a compact domain with the signed Radon measures of finite variation in* [119].

9.2 Distributed integer controls in space

This section considers the state vector approximation for the class of elliptic control problems (MIPELL) from Section 3.3. Here, the state vector approximation (c) follows straightforwardly and the most effort in proving (1.1) for (MIPELL) has been invested into (b). This section follows the arguments of the author in [79, Sect. 3 & 4.5].

Theorem 9.5 ([79, Thm 3.2])**.** *Let Assumption 3.7 hold. Let $\boldsymbol{\alpha} \in L^\infty(\Omega, \mathbb{R}^M)$ and the sequence $(\boldsymbol{\omega}^{(n)})_n \in L^\infty(\Omega, \mathbb{R}^M)$ satisfy the convergence*

$$\boldsymbol{\omega}^{(n)} \rightharpoonup^* \boldsymbol{\alpha}$$

in the space $L^\infty(\Omega, \mathbb{R}^M)$. Let S_R denote the control-to-state operator defined by $S_R(u, \beta) := A^{-1} \sum_{i=1}^M \beta_i f(u, v_i)$ for the controls $u \in \mathcal{U}$ and $\beta \in L^\infty(\Omega, \mathbb{R}^M)$. Let $y := S_R(u, \boldsymbol{\alpha})$ and $y^{(n)} := S_R(u, \boldsymbol{\omega}^{(n)})$ for $n \in \mathbb{N}$. Then, the convergence

$$y^{(n)} \to y$$

holds in the spaces $\mathcal{V}$ and $L^2(\Omega)$.

Proof. The convergence $\boldsymbol{\omega}^{(n)} \rightharpoonup^* \boldsymbol{\alpha}$ implies $\sum_{i=1}^M \boldsymbol{\omega}_i^{(n)} f(u, v_i) \rightharpoonup \sum_{i=1}^M \boldsymbol{\alpha}_i f(u, v_i)$ in $L^2(\Omega)$. The compact embedding $L^2(\Omega) \hookrightarrow^c \mathcal{V}^*$ asserted by Assumption 3.7 implies $\sum_{i=1}^M \boldsymbol{\omega}_i^{(n)} f(u, v_i) \to \sum_{i=1}^M \boldsymbol{\alpha}_i f(u, v_i)$ in $\mathcal{V}^*$, i.e. $Ay^{(n)} \to Ay$. By Assumption 3.7, the operator $A : \mathcal{V} \to \mathcal{V}^*$ is linear with bounded inverse, which yields the claim. □

Under an ellipticity condition and the assumption that the relaxed and binary controls $\boldsymbol{\alpha}$ and $\boldsymbol{\omega}$ are constant per grid cell, we prove the following a priori estimate.

Theorem 9.6 ([79, Thm 4.9])**.** *Let Assumption 3.7 hold. Let the ellipticity condition*

$$c_1 \|y\|_{\mathcal{V}}^2 \leq \langle Ay, y \rangle_{\mathcal{V}^*, \mathcal{V}}$$

hold for all $y \in \mathcal{V}$ for some $c_1 > 0$. Let $f_i \in L^2(\Omega)$ for $i \in \{1, \dots, M\}$. Then, there exists a constant $C > 0$ such that

$$\|y(\boldsymbol{\alpha}) - y(\boldsymbol{\omega})\|_{\mathcal{V}} \leq C d(\boldsymbol{\alpha}, \boldsymbol{\omega})$$

if d is the integrality gap of a rounding grid $\{\mathcal{T}_1, \dots, \mathcal{T}_N\}$ for Ω and all relaxed controls $\boldsymbol{\alpha}$ and binary controls $\boldsymbol{\omega}$ that are both a.e. constant per grid cell $\mathcal{T}_k$ and $y(\boldsymbol{\alpha})$ solving $Ay = \sum_{i=1}^M \boldsymbol{\alpha}_i f_i$ as well as $y(\boldsymbol{\omega})$ solving $Ay = \sum_{i=1}^M \boldsymbol{\omega}_i f_i$.

Proof. Let $z \coloneqq y(\boldsymbol{\alpha}) - y(\boldsymbol{\omega})$, $\boldsymbol{\phi} \coloneqq \boldsymbol{\alpha} - \boldsymbol{\omega}$. The ellipticity condition yields

$$c_1 \|z\|_{\mathcal{V}}^2 \leq \langle Az, z \rangle_{\mathcal{V}^*, \mathcal{V}} = \int_\Omega \sum_{i=1}^M \boldsymbol{\phi}_i(s) f_i(s) z(s) \, \mathrm{d}s. \tag{9.4}$$

We employ Lemma 7.4 to the term $|\int_\Omega \boldsymbol{\phi}_i f_i z|$ for all $i \in \{1, \dots, M\}$, i.e.

$$\left| \int_\Omega \boldsymbol{\phi}_i f_i z \right| \leq 2 d(\boldsymbol{\omega}, \boldsymbol{\alpha}) \|f_i z\|_{L^1}.$$

The triangle inequality and the ellipticity estimate give

$$\|z\|_{\mathcal{V}}^2 \leq d(\boldsymbol{\omega}, \boldsymbol{\alpha}) \frac{2M}{c_1} \max_{i \in \{1, \dots, M\}} \|f_i z\|_{L^1}.$$

The Cauchy–Schwarz inequality and the continuous embedding $\mathcal{V} \hookrightarrow L^2(\Omega)$ give

$$\begin{aligned} \|z\|_{\mathcal{V}}^2 &\leq d(\boldsymbol{\omega}, \boldsymbol{\alpha}) \frac{2M}{c_1} \max_{i \in \{1, \dots, M\}} \|f_i\|_{L^2} \|z\|_{L^2} \\ &\leq d(\boldsymbol{\omega}, \boldsymbol{\alpha}) \frac{2M c_2}{c_1} \max_{i \in \{1, \dots, M\}} \|f_i\|_{L^2} \|z\|_{\mathcal{V}} \end{aligned}$$

for some $c_2 > 0$. The choice $C \coloneqq \frac{2Mc_2}{c_1} \max_{i \in \{1, \dots, M\}} \|f_i\|_{L^2}$ closes the proof. □

9.3 Distributed integer controls in time and space

This section considers the state vector approximation for the evolution problems constraining (MIPEVO-TX). We recall the convexified IVP

$$\begin{aligned} \partial_t y + Ay &= -\sum_{i=1}^{M} \boldsymbol{\alpha}_i f^a(y, v_i) - f^b(y, u) \\ y(0) &= y_0 \\ y|_{\partial\Omega} &= 0 \end{aligned} \tag{3.4 revisited}$$

for a space- and time-dependent relaxed control $\boldsymbol{\alpha}$ in the setting specified in Assumptions 3.10, 3.11 and 3.13. We recall the state spaces

$$\mathcal{W} \coloneqq W((0,T)) = \left\{ y \in L^2((0,T), H_0^1(\Omega)) : \partial_t y \in L^2((0,T), H^{-1}(\Omega)) \right\}$$

and

$$\mathcal{Y} \coloneqq L^2((0,T), L^2(\Omega))$$

with the compact embedding $\mathcal{W} \hookrightarrow^c \mathcal{Y}$. Before proving the approximation result, we prove two preparatory lemmata. The first one states a certain interplay between vectors in Bochner and Lebesgue spaces, which we use several times.

Lemma 9.7. *Let $f \in L^2((0,T), L^2(\Omega))$ and $\boldsymbol{\omega} \in L^\infty(\Omega_T)$. Then,*

$$\boldsymbol{\omega} f \in L^2((0,T), L^2(\Omega)).$$

Proof. By virtue of Fubini's theorem, we can use the identification $L^2((0,T), L^2(\Omega)) \cong L^2(\Omega_T)$, see e.g. [60, Chap. 1.2.b]. Using Hölder's inequality, we have

$$\boldsymbol{\omega} f \in L^2(\Omega_T).$$

Employing the identification $L^2((0,T), L^2(\Omega)) \cong L^2(\Omega_T)$ again yields the claim. □

The proof of the state vector approximation will combine existence and uniqueness of the convexified state equation asserted in Theorem 3.12 with the weak approximation properties of the binary controls. The following lemma[7] considers weak convergence of the convexified summand separately.

Lemma 9.8. *Let $f : \mathcal{Y} \to L^2((0,T), L^2(\Omega))$ be continuous. Let the sequence $(y^{(n)})_n \subset \mathcal{Y}$ be convergent, $y^{(n)} \to y$ in $\mathcal{Y}$. Let $(\boldsymbol{\omega}^{(n)})_n \subset L^\infty(\Omega_T)$ be a sequence that satisfies $\boldsymbol{\omega}^{(n)} \rightharpoonup^* \boldsymbol{\alpha}$ for the limit $\boldsymbol{\alpha} \in L^\infty(\Omega_T)$. Then,*

$$\boldsymbol{\omega}^{(n)} f(y^{(n)}) \rightharpoonup \boldsymbol{\alpha} f(y)$$

in $L^2((0,T), L^2(\Omega))$.

[7] We acknowledge that the idea that the statement of Lemma 9.8 is the key to prove the state vector approximation for integer controls in Section 9.3 is due to Christian Meyer, TU Dortmund.

Proof. The continuity of $f : \mathcal{Y} \to L^2((0,T), L^2(\Omega))$ implies convergence of the sequence $(f(y^{(n)}))_n$ in $L^2((0,T), L^2(\Omega))$. Furthermore, we have weak* convergence of the sequence $(\boldsymbol{\omega}^{(n)})_n$ in $L^\infty(\Omega_T)$. Thus, Lemma 9.7 asserts that the sequence $(\boldsymbol{\omega}^{(n)} f(y^{(n)}))_n$ is bounded in $L^2((0,T), L^2(\Omega))$. Passing to a subsequence, we obtain a weak limit. The insertion of a suitable zero shows that the limit coincides with the product of the limits $f(y)$ and $\boldsymbol{\alpha}$. Passing to subsubsequences of subsequences, we obtain

$$\boldsymbol{\omega}^{(n)} f(y^{(n)}) \rightharpoonup \boldsymbol{\alpha} f(y)$$

in $L^2((0,T), L^2(\Omega))$. □

With these lemmata, we can prove the state vector approximation for the IVP (3.4).

Theorem 9.9. *Let Assumptions 3.10, 3.11 and 3.13 hold. Let $u \in L^q(\Omega_T)$. Let $(\boldsymbol{\omega}^{(n)})_n$ be a sequence of binary controls and $\boldsymbol{\alpha}$ be a relaxed control such that the convergence $\boldsymbol{\omega}^{(n)} \rightharpoonup^* \boldsymbol{\alpha}$ in $L^\infty(\Omega_T, \mathbb{R}^M)$ holds. For all $n \in \mathbb{N}$, let $y^{(n)}$ solve* (3.4) *for the controls u and $\boldsymbol{\omega}^{(n)}$. Let y solve* (3.4) *for the controls u and $\boldsymbol{\alpha}$. Then, the convergence*

$$y^{(n)} \rightharpoonup y \text{ in } \mathcal{W}$$

and the convergence

$$y^{(n)} \to y \text{ in } \mathcal{Y}$$

hold.

Proof. We abbreviate $z^{(n)} := (u, \boldsymbol{\omega}^{(n)})$ for $n \in \mathbb{N}$. The convergence $\boldsymbol{\omega}^{(n)} \rightharpoonup^* \boldsymbol{\alpha}$ implies the boundedness $\|z^{(n)}\|_{L^q(\Omega_T, \mathbb{R}^{1+M})} \leq C_1$ for some $C_1 > 0$ and consequently, the boundedness $\|y^{(n)}\|_{L^\infty(\Omega_T)} \leq C_2$ for some $C_2 > 0$ by Theorem 3.12. For $(y, u, \alpha) \in \mathbb{R}^{2+M}$, we abbreviate $f(y,z) := \sum_{i=1}^M \alpha_i f_i^a(y, v_i) + f^b(y,u)$. Next, we show boundedness of the sequence $(y^{(n)})_n$ in $\mathcal{W}$ to obtain weakly convergent subsequences.

Boundedness of $(y^{(n)})_n \subset \mathcal{W}$: By Assumption 3.11, we can use the identity $f(y,z) = f(0,z) + \int_0^1 \partial_y f(\theta y, z) y \,\mathrm{d}\theta$ to estimate

$$\begin{aligned}
&\|f(y^{(n)}, z^{(n)})\|_{L^q(\Omega_T)}^q \\
&\qquad \leq 2^{q-1} \int_{\Omega_T} |f(0, z^{(n)})|^q + \left(\int_0^1 |\partial_y f(\theta y^{(n)}, z^{(n)})| \,\mathrm{d}\theta \right)^q |y^{(n)}|^q \,\mathrm{d}(t,x).
\end{aligned}$$

We bound the first summand with the constants M_1, $m_1 > 0$ from Assumption 3.11 as

$$\int_{\Omega_T} |f(0, z^{(n)})|^q \,\mathrm{d}(t,x) \leq 2^{q-1}(2^q M_1^q \lambda(\Omega_T) + m_1^q \|u\|_{L^q}^q).$$

Using the same constants, the additional constant $|C_0| > 0$ and the non-decreasing function $\eta : \mathbb{R}_+ \to \mathbb{R}_+$, we estimate the second summand as

$$\begin{aligned}
&\int_{\Omega_T} \left(\int_0^1 |\partial_y f(\theta y^{(n)}, z^{(n)})| \,\mathrm{d}\theta \right)^q |y^{(n)}|^q \,\mathrm{d}(t,x) \\
&\qquad \leq \int_{\Omega_T} \left(\int_0^1 \max\{|C_0|, (2M_1 + m_1|u|)\eta(|y^{(n)}|)\} \,\mathrm{d}\theta \right)^q C_2^q \,\mathrm{d}(t,x) \\
&\qquad \leq \left(2^{q-1} \left(2^q M_1^q \lambda(\Omega_T) + m_1^q \|u\|_{L^q}^q\right) \eta(C_2)^q + |C_0|^q \right) C_2^q.
\end{aligned}$$

We continue with a bootstrapping argument. For all $n \in \mathbb{N}$, we consider the IVPs

$$\partial_t w^{(n)} + A w^{(n)} = \varphi^{(n)}, \quad w^{(n)}(0) = y_0, \tag{9.5}$$

in which we have defined the vectors $\varphi^{(n)} := -f(y^{(n)}, z^{(n)})$. Importantly, the sequence $(\varphi^{(n)})_n$ is bounded in $L^q(\Omega_T)$ by the considerations above. The boundedness and the uniqueness of the solutions to (3.4) give the identities $w^{(n)} = y^{(n)}$ for all $n \in \mathbb{N}$. Similar to the proof of Lemma B.48 or the argument in [26, Chap. XVIII, Sect. §3.3.2], we obtain

$$\begin{aligned} &\|w^{(n)}\|_{L^2((0,T),H_0^1(\Omega))} \\ &\quad \leq C_3 \left(\|\varphi^{(n)}\|_{L^2((0,T),H^{-1}(\Omega))} + \|y_0\|_{L^2(\Omega)} \right) \\ &\quad \leq C_3 \left(C_4 \|\varphi^{(n)}\|_{L^q(\Omega_T)} + \|y_0\|_{L^2(\Omega)} \right) \end{aligned} \tag{9.6}$$

for some constants $C_3 > 0$ and $C_4 > 0$. We estimate the derivative as

$$\begin{aligned} \|\partial_t w^{(n)}\|_{L^2((0,T),H^{-1}(\Omega))} &\underset{(9.5)}{\leq} \|A w^{(n)}\|_{L^2((0,T),H^{-1}(\Omega))} + \|\varphi^{(n)}\|_{L^2((0,T),H^{-1}(\Omega))} \\ &\underset{\text{Ass. 3.10}}{\leq} M \|w^{(n)}\|_{L^2((0,T),H_0^1(\Omega))} + C_4 \|\varphi^{(n)}\|_{L^q(\Omega_T)} \\ &\underset{(9.6)}{\leq} M \left(C_4 \|\varphi^{(n)}\|_{L^q(\Omega_T)} + \|y_0\|_{L^2(\Omega)} \right) + C_4 \|\varphi^{(n)}\|_{L^q(\Omega_T)}. \end{aligned}$$

Finally, we have asserted that $(y^{(n)})_n$ is a bounded subset of $\mathcal{W}$.

The convergence $y^{(n)} \to y$ in $\mathcal{Y}$ The boundedness of the sequence $(y^{(n)})_n \subset \mathcal{W}$ and the reflexivity of $\mathcal{W}$, which follows from Theorem B.27 and the reflexivity of $H_0^1(\Omega)$, implies the existence of a weakly convergent subsequence $(y^{(n_k)})_k$, i.e.

$$y^{(n_k)} \rightharpoonup \bar{y}$$

in $\mathcal{W}$ for some $\bar{y} \in \mathcal{W}$ and by compactness of the embedding $\mathcal{W} \hookrightarrow^c \mathcal{Y}$, we obtain

$$y^{(n_k)} \to \bar{y}$$

in $\mathcal{Y}$. Let $\varphi \in L^2((0,T), H_0^1(\Omega))$. Then, the equality

$$\langle (\partial_t + A) y^{(n_k)}, \varphi \rangle = - \left\langle \sum_{i=1}^{M} \boldsymbol{\omega}_i^{(n_k)} f_i^a(y^{(n_k)}) + f^b(y^{(n_k)}, u), \varphi \right\rangle$$

holds for all n_k and the pairing $\langle \cdot, \cdot \rangle$ that puts $L^2((0,T), H_0^1(\Omega))$ and $L^2((0,T), H^{-1}(\Omega))$ in duality. The convergence $y^{(n_k)} \rightharpoonup \bar{y}$ in $\mathcal{W}$ implies $\partial_t y^{(n_k)} \rightharpoonup \partial_t \bar{y}$. Furthermore, the mapping $A : L^2((0,T), H_0^1(\Omega)) \to L^2((0,T), H^{-1}(\Omega))$ is a norm-norm continuous linear operator between Hilbert spaces. Consequently, it is also weak-weak continuous and we obtain $A y^{(n_k)} \rightharpoonup A\bar{y}$. Consequently, we can pass to the limit on the left side. Using Assumption 3.13 and Lemma 9.8, we can do this on the right side as well and obtain

$$\langle (\partial_t + A) \bar{y}, \varphi \rangle = - \left\langle \sum_{i=1}^{M} \boldsymbol{\alpha}_i f_i^a(\bar{y}) + f^b(\bar{y}, u), \varphi \right\rangle,$$

which means that $\bar{y}$ is a weak solution of (3.4) for the controls u and $\boldsymbol{\alpha}$. As the solution of (3.4) is unique, we obtain $y = \bar{y}$. Passing to subsubsequences, this argument applies for every subsequence and we obtain

$$y^{(n)} \rightharpoonup y \text{ in } \mathcal{W} \text{ and } y^{(n)} \to y \text{ in } \mathcal{Y}.$$

□

9.4 Convolution operators with fixed kernels

Finally, we briefly consider the state vector approximation for state equations of the form

$$y = K\boldsymbol{v}, \tag{9.7}$$

in which K is a convolution operator induced by a fixed kernel function. Let $\Omega \subset \mathbb{R}^d$ be a bounded domain for $d \in \mathbb{N}$. Let $k \in L^1(\mathbb{R}^d)$. For the remainder of this section, let

$$(Kf)(x) \coloneqq (k * f)(x) = \int_\Omega k(x-s)f(s)\,\mathrm{d}s \tag{9.8}$$

for $f \in L^2(\Omega)$. Importantly for our considerations, the operator K is compact, which is asserted in the following proposition.

Proposition 9.10 ([107, Thm 3.1.17])**.** *Let $\Omega \subset \mathbb{R}^d$ be a bounded domain. Let $k \in L^1(\mathbb{R}^d)$. Then, the operator $K : L^2(\Omega) \to L^2(\Omega)$ is compact (see Definition B.4).* ■

Consequently, we can prove implication (c) for the problem class (MIOCP) if the control-to-state operator is defined by a convolution operation.

Theorem 9.11. *Let $\Omega \subset \mathbb{R}^d$ be a bounded domain. Let $\{v_1, \dots, v_M\} \subset \mathbb{R}^{n_v}$ for some $n_v \in \mathbb{N}$. Let $k \in L^1(\mathbb{R}^d)$. Let $\boldsymbol{\alpha}$ be a relaxed control and let $(\boldsymbol{\omega}^{(n)})_n$ be a sequence of binary controls such that*

$$\boldsymbol{\omega}^{(n)} \rightharpoonup^* \boldsymbol{\alpha} \text{ in } L^\infty(\Omega, \mathbb{R}^M).$$

Then, the vectors $y^{(n)} \coloneqq K\left(\sum_{i=1}^M \boldsymbol{\omega}_i^{(n)} v_i\right)$ satisfy

$$y^{(n)} \to y \text{ in } L^2(\Omega)$$

for the limit $y = K\left(\sum_{i=1}^M \boldsymbol{\alpha}_i v_i\right)$.

Proof. The convergence $\boldsymbol{\omega}^{(n)} \rightharpoonup^* \boldsymbol{\alpha}$ implies $\boldsymbol{\omega}_i^{(n)} v_i \rightharpoonup^* \sum_{i=1}^M \boldsymbol{\alpha}_i v_i$ in $L^\infty(\Omega, \mathbb{R}^{n_v})$. Then, the claim follows from Proposition 9.10. □

For the one-dimensional case, we are able to state an a priori estimate on the state vector approximation error

$$\|K\boldsymbol{v} - K\boldsymbol{v}^{(h)}\|_{L^2},$$

where $\boldsymbol{v} \coloneqq \sum_{i=1}^{M} \boldsymbol{\alpha}_i v_i$ and $\boldsymbol{v}^{(h)} \coloneqq \sum_{i=1}^{M} \boldsymbol{\omega}_i^{(h)} v_i$ and $\boldsymbol{\omega}^{(h)}$ is computed by a rounding algorithm from $\boldsymbol{\alpha}$. The estimate is linear in the integrality gap $d(\boldsymbol{\omega}^{(h)}, \boldsymbol{\alpha})$. The proof hinges on the integration by parts formula, which causes the restriction to the one-dimensional case and an additional differentiability assumption on the kernel function.

Theorem 9.12. *Let $C > 0$ and t_0, $t_f \in \mathbb{R}$ with $t_0 < t_f$. Let $\boldsymbol{\alpha}$ and $\boldsymbol{\omega}$ be a relaxed and a binary control in $L^\infty((t_0, t_f))$. Let $k \in W^{1,\infty}(\mathbb{R})$. Then, the estimate*

$$\left\| K\left(\sum_{i=1}^{M} \boldsymbol{\alpha}_i v_i\right) - K\left(\sum_{i=1}^{M} \boldsymbol{\omega}_i v_i\right) \right\|_{L^2} \leq L d(\boldsymbol{\alpha}, \boldsymbol{\omega})$$

holds with $L \coloneqq M \max_{i \in \{1,\dots,M\}} \|v_i\|_\infty \max\{1, t_f - t_0\} \|k\|_{W^{1,\infty}} \sqrt{t_f - t_0}$.

Proof. We use the notation $\boldsymbol{v}$ and $\boldsymbol{v}^{(h)}$ from above. The linearity of the convolution gives $k * \boldsymbol{v} - k * \boldsymbol{v}^{(h)} = k * (\boldsymbol{v} - \boldsymbol{v}^{(h)})$. The convolution may be interpreted pointwise here as the convolution of two functions from conjugate L^p spaces is continuous, which follows from Hölder's inequality and passing to the uniform limit in an approximating sequence of continuous functions. Let $t \in [t_0, t_f]$. We employ integration by parts and the continuous embedding $W^{1,\infty}((t_0 - t_f, t_f - t_0)) \hookrightarrow C([t_0 - t_f, t_f - t_0])$ to obtain

$$\begin{aligned}
&(k * (\boldsymbol{v} - \boldsymbol{v}^{(h)}))(t) \\
&\quad = \int_{t_0}^{t_f} k(t-s)(\boldsymbol{v} - \boldsymbol{v}^{(h)})(s) \,\mathrm{d}s \\
&\quad = k(t - t_f) \int_{t_0}^{t_f} (\boldsymbol{v} - \boldsymbol{v}^{(h)})(s) \,\mathrm{d}s - \int_{t_0}^{t_f} (-k')(t-s) \int_{t_0}^{s} (\boldsymbol{v} - \boldsymbol{v}^{(h)})(\tau) \,\mathrm{d}\tau \,\mathrm{d}s.
\end{aligned}$$

Using this equality, we are able to estimate

$$\begin{aligned}
&\left|(k * (\boldsymbol{v} - \boldsymbol{v}^{(h)}))(t)\right| \\
&\quad \leq \|k\|_{L^\infty} \left| \sum_{i=1}^{M} v_i \int_{t_0}^{t_f} \boldsymbol{\alpha}_i(s) - \boldsymbol{\omega}_i(s) \,\mathrm{d}s \right| \\
&\qquad + \|k'\|_{L^\infty} \left| \sum_{i=1}^{M} v_i \int_{t_0}^{t_f} \int_{t_0}^{s} \boldsymbol{\alpha}_i(\tau) - \boldsymbol{\omega}_i(\tau) \,\mathrm{d}\tau \,\mathrm{d}s \right| \\
&\quad \leq \sum_{i=1}^{M} |v_i| \|k\|_{L^\infty} d(\boldsymbol{\omega}, \boldsymbol{\alpha}) + \|k'\|_{L^\infty} \sum_{i=1}^{M} |v_i| (t_f - t_0) d(\boldsymbol{\omega}, \boldsymbol{\alpha}) \\
&\quad \leq M \max_{i \in \{1,\dots,M\}} \|v_i\|_\infty \max\{1, t_f - t_0\} \|k\|_{W^{1,\infty}} d(\boldsymbol{\omega}, \boldsymbol{\alpha}).
\end{aligned}$$

The first inequality follows from Hölder's inequality, the second from the triangle inequality and again Hölder's inequality and the third from the definition of $\|\cdot\|_{W^{1,\infty}}$. Inserting this estimate into the definition of the L^2-norm concludes the proof. □

Part III

Applications and computational results

Chapter 10

Computational examples

In this chapter, we demonstrate our findings computationally, which is contribution (e). We begin by demonstrating the approximation arguments in (1.1) for a semilinear evolution equation with a discrete control input that is distributed in the time domain in Section 10.1. Section 10.1 is based on [80, Sect. 4]. In Section 10.2, we demonstrate (1.1) and the optimality relationship discussed in Chapter 4 for a signal processing problem, in which the discrete control input is also distributed in the time domain. In Section 10.3, we demonstrate (1.1), in particular (b), for a discrete control that is distributed in two dimensions using the Dirichlet Laplacian as differential operator in the state equation. Section 10.3 is based on [79, Sect. 6]. We close with a little excursion in Section 10.4, in which we briefly summarize the analysis yielding the necessary compactness to obtain (c) for fractional powers of the Dirichlet Laplacian, which in turn allows us to obtain and demonstrate (1.1) in this case as well.

10.1 State vector approximation for a transport equation

This section serves to demonstrate the chain of approximation arguments

$$h \to 0 \quad \underset{(a)}{\Rightarrow} \quad d_{\mathrm{1D}}(\boldsymbol{\omega}^{(h)}, \boldsymbol{\alpha}) \to 0 \quad \underset{(b)}{\Rightarrow} \quad \boldsymbol{\omega}^{(h)} \rightharpoonup \boldsymbol{\alpha} \quad \underset{(c)}{\Rightarrow} \quad y(\boldsymbol{\omega}^{(h)}) \to y(\boldsymbol{\alpha}) \quad \text{(1.1 revisited)}$$

for an IVP that is governed by a strongly continuous semigroup with time-dependent relaxed and binary controls in the absence of differentiability of the state vector trajectory and derived quantities. We follow the author's considerations in [80, Sect. 3]. Specifically, we validate Theorem 9.3 computationally for the IVP

$$\begin{aligned} \partial_t y(t) + \partial_x y(t) &= \boldsymbol{\alpha}_1(t) f_1(t) + \boldsymbol{\alpha}_2(t) f_2(t) \\ y(0) &\equiv 0.5. \end{aligned} \tag{10.1}$$

We assume a one-dimensional domain $\Omega = (\ell, r)$, a constant influx of 0 from the left boundary of the domain and do not impose any condition at the right boundary. We notice that the operator ∂_x generates the right translation semigroup, see [87,

Example 2.9]. Importantly for this example, the translation semigroup does not induce a smoothing of the state vector trajectory like e.g. the heat semigroup does. Regarding the right hand side of the PDE, we set the function f_1 to the product of a Weierstraß function that is nowhere differentiable in $[0,T]$, see Example 7.2 for details, and an indicator function for some part of the spatial domain. We set $f_2(t) \equiv 0$. As the arguments in (1.1) do not require optimality of the approximated pair $(y, \boldsymbol{\alpha})$, we can choose $\boldsymbol{\alpha}$ freely and use $\boldsymbol{\alpha}_1 = \boldsymbol{\alpha}_2 \equiv 0.5$.

We discretize the time horizon $[0, 10]$ and the domain Ω into 4096 intervals each and solve the IVP numerically by means of the Lax-Friedrichs scheme, see [72] for the details and further reading on numerics for hyperbolic equations.

To compute the weak approximants of $\boldsymbol{\alpha}$, we employ the algorithm (SUR), which produces the chattering we have already observed in the right column of Fig. 2.1. We consider the binary controls $\boldsymbol{\omega}^{(1)}, \dots, \boldsymbol{\omega}^{(6)}$ that are computed by (SUR) on rounding grids consisting of $N^{(1)} = 128, \dots, N^{(6)} = 4096$ intervals and the corresponding solutions $y^{(1)}, \dots, y^{(6)}$ of the IVP (10.1). Furthermore, we calculate the relative error

$$\epsilon^{(n)} := \frac{\sup_{t\in[0,T]} \|y^{(n)}(t) - y(t)\|_{L^1}}{\sup_{t\in[0,T]} \|y(t)\|_{L^1}}$$

for $n = 1, \dots, 6$. As we cannot evaluate the Weierstraß function exactly, we have approximated it using $k = 1$, 10, 100 summands of its series representation. We observe that $\epsilon^{(n)}$ tends to 0 for the three choices of k in a similar fashion. If only one or two summands of the cosine series are included, i.e. the smoothness is *highest*, the convergence is a little faster than for the other cases. In numbers, we have $\epsilon^{(6)} = 1.6642 \cdot 10^{-3}$ for $k = 1$, $\epsilon^{(6)} = 2.1519 \cdot 10^{-3}$ for $k = 10$ and $\epsilon^{(6)} = 2.1521 \cdot 10^{-3}$ for $k = 100$. We have visualized the behavior of the sequence $(\epsilon^{(n)})_n$ in Fig. 10.1. To demonstrate the non-differentiability in the right hand side, we have plotted the

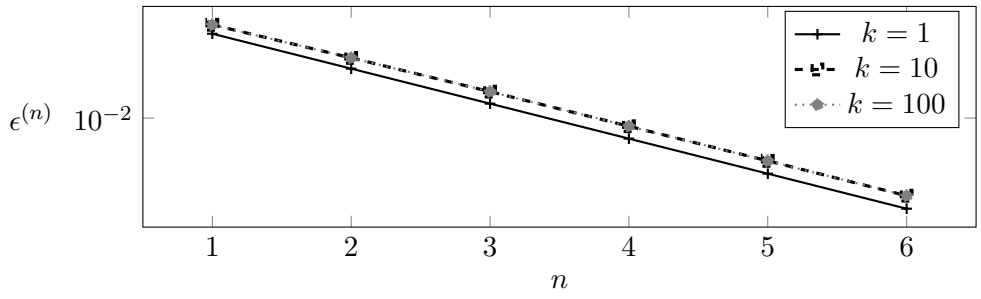


Figure 10.1: Decline of the relative state vector approximation error $\epsilon^{(n)}$ for refined grids indexed by $n = 1, \dots, 6$. The figure has been published in [80] as Figure 2.

approximants of the Weierstraß function f_1 in Fig. 10.2.

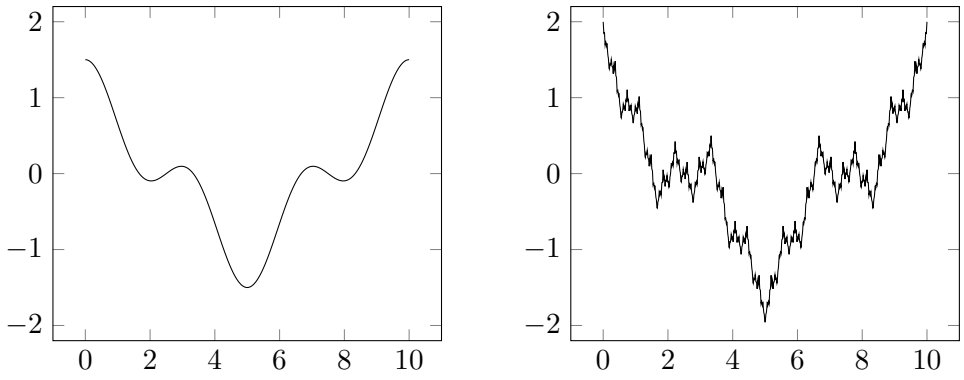


Figure 10.2: Approximants of f_1 for two summands (left) and ten summands (right). The figure has been published in [80] as Figure 3.

10.2 Filtered signal approximation

This section demonstrates our chain of approximation arguments for the approximation of a given filtered target signal. In particular, we also demonstrate the optimality principle

$$y(\boldsymbol{\omega}^{(h)}) \to y^*(\boldsymbol{\alpha}^*) \underset{(d)}{\Rightarrow} \min_{(y,\boldsymbol{\alpha})\in\mathcal{F}_{(\mathrm{RC})}} J(y) = \inf_{(y,\boldsymbol{\omega})\in\mathcal{F}_{(\mathrm{BC}_0)}} J(y)$$

for minimizers $(y^*, \boldsymbol{\alpha}^*) \in \arg\min\{J(y) : (y, \boldsymbol{\alpha}) \in \mathcal{F}_{(\mathrm{RC})}\}$ of the relaxation. We restrict ourselves to a one-dimensional setting and consider the following optimization problem:

$$\begin{aligned} &\min_{\boldsymbol{v}} J(\boldsymbol{v}) = \frac{1}{2}\int_{t_0}^{t_f} ((k * \boldsymbol{v})(t) - f(t))^2 \,\mathrm{d}t \\ &\text{s.t. } \boldsymbol{v} \in L^2((t_0, t_f)) \\ &\qquad \boldsymbol{v}(t) \in \{v_1, \ldots, v_M\} \subset \mathbb{R} \text{ for a.e. } t \in (t_0, t_f) \end{aligned} \tag{MIOCP}$$

Here, we assume that $k \in L^2((t_0, t_f))$ is a fixed kernel function and $f \in L^2((t_0, t_f))$. Buchheim et al. have treated (MIOCP) successfully following the methodology

1. discretize (MIOCP) into a finite-dimensional Integer Program (IP),
2. solve the discretized problem with finite-dimensional IP-techniques

in [16] by means of a branch-and-bound algorithm for convex quadratic IPs, which exhibits high computational demand for fine discretizations. We employ our approximation arguments and the algorithm (SUR) to demonstrate that we can benefit from the infinite-dimensional structure of the problem and following the procedure

1. reformulate (MIOCP) into (BC_0) and relax (BC_0) into (RC),
2. discretize and solve the (RC),
3. compute roundings from the solution of (RC).

The reformulation reads

$$\begin{aligned} \min_{y,\boldsymbol{\omega}} J(y) &= \frac{1}{2}\|y-f\|_{L^2}^2 \\ & \boldsymbol{\omega} \in L^2((t_0,t_f),\mathbb{R}^M) \\ \text{s.t. } & y = k * \textstyle\sum_{i=1}^M \boldsymbol{\omega}_i v_i \\ & \boldsymbol{\omega}(t) \in \{0,1\}^M \text{ and } \textstyle\sum_{i=1}^M \boldsymbol{\omega}_i(t) = 1 \text{ for a.a. } t \in (t_0,t_f) \end{aligned} \tag{BC$_0$}$$

and is relaxed to

$$\begin{aligned} \min_{y,\boldsymbol{\alpha}} J(y) &= \frac{1}{2}\|y-f\|_{L^2}^2 \\ & \boldsymbol{\alpha} \in L^2((t_0,t_f),\mathbb{R}^M) \\ \text{s.t. } & y = k * \textstyle\sum_{i=1}^M \boldsymbol{\omega}_i v_i \\ & \boldsymbol{\alpha}(t) \in [0,1]^M \text{ and } \textstyle\sum_{i=1}^M \boldsymbol{\alpha}_i(t) = 1 \text{ for a.a. } t \in (t_0,t_f). \end{aligned} \tag{RC}$$

The state equation $y = k * \boldsymbol{v}$ with $\boldsymbol{v} = \sum_{i=1}^M \boldsymbol{\omega}_i v_i$ in (BC$_0$) and $\boldsymbol{v} = \sum_{i=1}^M \boldsymbol{\alpha}_i v_i$ in (RC) allows to apply Theorem 9.11, which yields (c), and Corollary 4.5, which yields (d).

We consider the example from [16], which stems from Filtered Approximation in electronics. We introduce a function

$$\kappa(t) \coloneqq A\left(1 - \sqrt{2}\exp\left(-\frac{\omega_0 t}{\sqrt{2}}\right)\cos\left(\frac{\omega_0 t}{\sqrt{2}} - \frac{\pi}{4}\right)\right)$$

with fixed parameter values $\omega_0 = \pi$ and $A = 0.1$ and define the convolution kernel for (MIOCP), (BC$_0$) and (RC) as

$$k(t) \coloneqq \begin{cases} (-\kappa)'(t) & t \geq 0, \\ 0 & \text{else}, \end{cases}$$

which yields $(k * \boldsymbol{v})(t) = \int_{t_0}^t k(t-\tau)\boldsymbol{v}(\tau)\,\mathrm{d}\tau$ for $\boldsymbol{v} \in L^2((t_0,t_f))$. We use $t_0 = -1$ and $t_f = 1$. Regarding the target function f, we set $f(t) \coloneqq 0.2\cos(2\pi t)$. The feasible realizations for $\boldsymbol{v}$ are $v_L = v_1 = -1$, $v_2 = 0$ and $v_U = v_3 = 1$. We discretize the resulting relaxation (RC) for the equivalent controls $\boldsymbol{v} = \sum_{i=1}^M \boldsymbol{\alpha}_i v_i$, $\boldsymbol{v} \in [v_L, v_U]$ with piecewise constant ansatz functions for $\boldsymbol{v}$ on an equidistant grid. Then, we optimize using `scipy.least_squares`, i.e. SciPy's *Trust Region Reflective* code, see [63]. The code is executed with the `tr_solver='exact'` option to avoid regularization in the solver as regularization terms are usually only weakly lower semi-continuous and not weakly continuous, which would prohibit the convergence to the minimal value of the relaxation, see also Corollary 4.6 and the considerations in Section 10.3. The algorithm terminated when the norm of the gradient fell below 10^{-10}.

We compute convex coefficient functions $\boldsymbol{\alpha}$ from $\boldsymbol{v}$ such that $\sum_{i=1}^M \boldsymbol{\alpha}_i v_i = \boldsymbol{v}$. As this may be non-unique, we have chosen the most-intuitive convex combination from our point of view: for $t \in [t_0,t_f]$, we compute $\boldsymbol{\alpha}(t)$ such that $\boldsymbol{v}(t)$ interpolates between its two neighboring points in $\{v_1,\dots,v_M\}$, i.e. we select i such that $v_i \leq \boldsymbol{v}(t) \leq v_{i+1}$ and set

$$\boldsymbol{\alpha}_i(t) \coloneqq \frac{v_{i+1} - \boldsymbol{v}(t)}{v_{i+1} - v_i},$$

$\boldsymbol{\alpha}_{i+1}(t) \coloneqq 1 - \boldsymbol{\alpha}_i(t)$ and $\boldsymbol{\alpha}_j(t) \coloneqq 0$ for $j \notin \{i, i+1\}$. Then, we apply (SUR) on a sequence of successively refined grids until the rounding grid coincides with the grid discretizing the state equation.

For 256 intervals discretizing (t_0, t_f), we have visualized the results in Fig. 10.3. The images in the left column show $y(\boldsymbol{\alpha}) = k * \boldsymbol{v}(\boldsymbol{\alpha}) = k * \left(\sum_{i=1}^{M} \boldsymbol{\alpha}_i v_i\right)$ and $y(\boldsymbol{\omega}) = k * \boldsymbol{v}(\boldsymbol{\omega}) = k * \left(\sum_{i=1}^{M} \boldsymbol{\omega}_i v_i\right)$ against the tracked function f for $\boldsymbol{\omega}$ being computed for $N = 4$, $N = 32$ and $N = 256$ rounding intervals. The convergence is clearly visible. The right column shows $\boldsymbol{v}(\boldsymbol{\alpha})$ and $\boldsymbol{v}(\boldsymbol{\omega})$ for the rounding grids consisting of $N = 4$, $N = 32$ and $N = 256$ intervals.

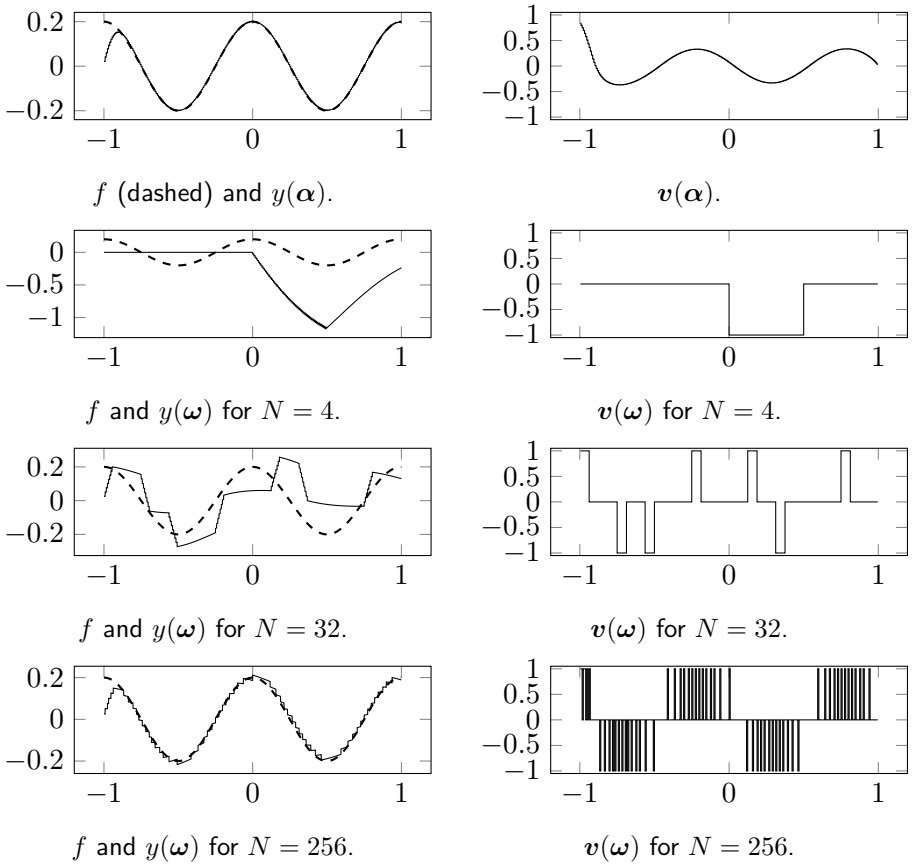


Figure 10.3: Relaxed control and state (top) and their SUR approximants for $N = 4$, $N = 32$ and $N = 256$ (rows two to four).

To close this section, we look quickly into the execution times of the code on a laptop computer equipped with a Intel(R) Core(TM) i7-6820 CPU clocked at 2.70 GHz. The main part of the computational costs is caused by the solution of (RC). The costs for the execution of (SUR) are negligible. The execution times for 2^i intervals, $i \in \{7, \ldots, 12\}$, are tabulated in Table 10.2.

Table 10.1: Relative differences between $J(y(\boldsymbol{\alpha}))$ and $J(y(\boldsymbol{\omega}))$.

N	$\frac{J(y(\boldsymbol{\omega}))-J(y(\boldsymbol{\alpha}))}{J(y(\boldsymbol{\alpha}))}$
2	4.8069×10^{1}
4	5.9671×10^{2}
8	4.6880×10^{2}
16	1.0251×10^{2}
32	3.8486×10^{1}
64	4.6024×10^{0}
128	1.6202×10^{0}
256	5.0305×10^{-1}
512	2.0503×10^{-1}
1024	2.0066×10^{-2}
2048	5.8747×10^{-3}
4096	1.4919×10^{-3}

Table 10.2: Execution times of the solution of (RC) for N intervals discretizing (t_0, t_f).

N	Time to solve (RC)
128	1.8104×10^{1} s
256	5.1509×10^{1} s
512	1.1827×10^{2} s
1024	3.2222×10^{2} s
2048	1.5094×10^{3} s
4096	9.2215×10^{3} s

10.3 Multi-dimensional elliptic systems

This section follows the work of the author in [79, Sect. 6]. We demonstrate the findings from Sections 7.2 and 9.2 computationally. The meshes and PDE solutions have been implemented using the FEniCS toolbox [3]. We consider the Dirichlet Laplacian as elliptic operator, which satisfies our assumptions, see Example 10.1 below.

Example 10.1. *We consider the Dirichlet Laplacian on the unit square* $\bar{\Omega} = [0,1]^2$ *such that the constraints of* (RC) *read*

$$\begin{aligned} -\Delta y &= \textstyle\sum_{i=1}^{M} \boldsymbol{\alpha}_i v_i \\ y|_{\partial\Omega} &= 0 \\ \boldsymbol{\alpha}(x) &\in \operatorname{conv} \mathbb{S}^M \ \textit{for a.a. } x \in \Omega. \end{aligned}$$

This setup is well-posed by Proposition 3.8.

In all experiments in this section, we compute the binary control approximants on uniformly refined uniform grids of squares. In each refinement, the side lengths of the squares are halved, which quadruples the number of squares. Section 10.3.1 briefly addresses the question, which progression through the grid cells to choose in implementations of (SUR-GEN). In Section 10.3.2, the multi-dimensional variant of (SUR) and the induced convergence properties are demonstrated. Afterwards, we approach a tracking-type MIOCP that is constrained by the Dirichlet Laplacian in Section 10.3.3, and also demonstrate the limitations mentioned in Section 4.2.

10.3.1 Cell progression for the SUR algorithm

Admissible sequences of refined rounding grids arise from uniform grid refinements. Assuming that the binary controls are computed with (SUR) or (SUR-VC) for a fixed relaxed control on an admissible sequence of refined rounding grids, Theorem 7.9 asserts weak

convergence regardless of the indexing of the grid cells. An earlier proof of the author for the weak convergence, see [81], demanded that the indexing of the grid cells is preserved along the grid refinements in a way, which is e.g. satisfied by approximants of iterates of space-filling curves, e.g. the Hilbert curve, see the facsimile in Fig. 10.4 of Hilbert's figure in [53]. Thus, we check if these orderings work well in practical implementations.

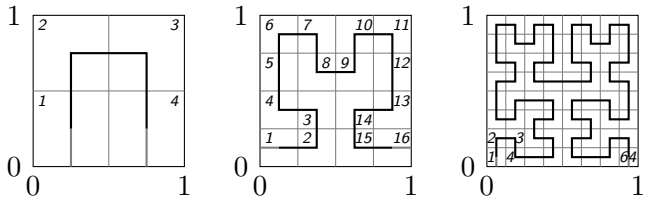


Figure 10.4: The first three Hilbert curve iterates on $\bar{\Omega} = [0,1]^2$. The additional extension to the boundary is marked gray. The induced grid cells are circumscribed by the gray lines. The grid cell indexing along the iterates is indicated by the small numbers inside the cells. The figure has been published as Figure 1 in [79].

To this end, (SUR) is executed 21 times on eight successive uniform refinements: once along the indexing induced by the Hilbert curve approximants and 20 times along random permutations of this indexing. A grayscale image of Hilbert serves as relaxed control. The resulting weak convergence of the binary controls to the relaxed control is visualized in Fig. 10.5 for the Hilbert curve-induced indexing. We also visualize the corresponding state vector approximation errors in Fig. 10.6. We perceive a linear decrease of the state vector approximation error for the random permutations visually. The Hilbert curve-induced indexing yields a considerably faster decrease than the random indexing.

Figure 10.5: Weak approximants computed with (SUR) for a grayscale image of David Hilbert along the order defined by the 1st, 3rd, 5th, 7th and 9th Hilbert curve approximant. The figure has been published as Figure 2 in [79].

The computational results in the subsequent sections are produced by executing (SUR) along the Hilbert curve-induced cell indexing.

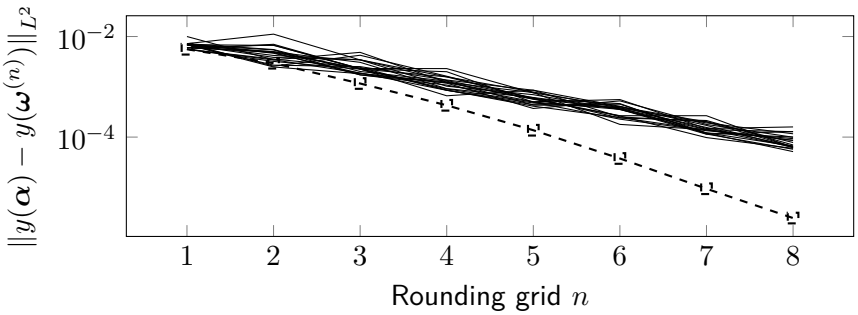


Figure 10.6: State approximation error for (SUR) on uniformly refined grids along Hilbert curve-induced cell indexing (dashed) and along 20 randomly permuted indexings (black, solid). The figure has been published as Figure 3 in [79].

10.3.2 Illustration of the approximation arguments

For the right hand side of the BVP in Example 10.1, let $M = 3$,

$$v_1 = 0, \quad v_2 = 2.1\chi_{B_{0.1}((0.172,0.3))}, \quad v_3 = 2\chi_{B_{0.123}((0.7,0.7))}$$

and let $\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \boldsymbol{\alpha}_3 : \Omega \to [0,1]$ with $\sum_{i=1}^{3} \boldsymbol{\alpha}_i = 1$ in Ω be given as visualized in Fig. 10.7a. We visualize the resulting integrality gap from the application of (SUR) and the expected bounds in Fig. 10.7b. We demonstrate Theorem 9.5 and Theorem 7.9, i.e. the implications (b) and (c), with the aforementioned $\boldsymbol{\alpha}_i$ and v_i in Fig. 10.8.

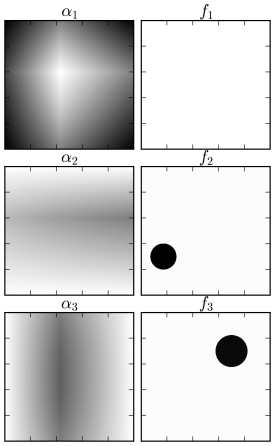
The corresponding relative approximation errors of the state vectors in the L^2- and the H^1-norm are visualized in Fig. 10.9. Interestingly, the human eyes and brain are able to sense the weak convergence $\boldsymbol{v}^{(n)} := \sum_i \boldsymbol{\omega}_i^{(n)} v_i \rightharpoonup \sum_i \boldsymbol{\alpha}_i v_i =: \boldsymbol{v}$ in Figs. 10.5 and 10.8. The output of (SUR) and (SUR-VC) is similar to the one of dithering techniques from computer graphics, which have been used to display grayscale images with coarsely quantized gray colors as e.g. performed by the Floyd-Steinberg algorithm [113].

10.3.3 Approximating the solution of an elliptic control problem

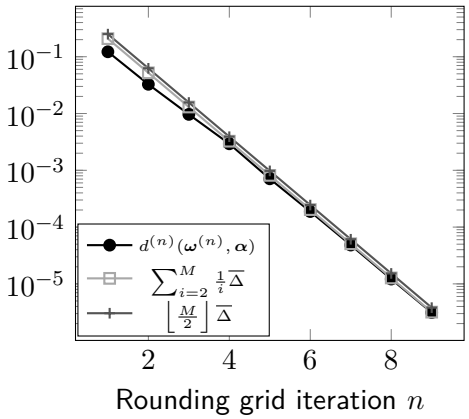
We consider the optimal control problem

$$\begin{aligned} \min_{y,\boldsymbol{v}} \ & \tfrac{1}{2}\|y - y_d\|_{L^2}^2 + \tfrac{\gamma}{2}\|\boldsymbol{v}\|_{L^2}^2 \\ \text{s.t.} \ & -\Delta_D y = \boldsymbol{v} \\ & \boldsymbol{v}(x) \in \{v_1, \ldots, v_M\} \subset \mathbb{R} \text{ for a.a. } x \in \Omega \end{aligned} \tag{MIOCP}$$

with $v_1 < \ldots < v_M$, which is similar to the model problem (1.1) in [22] by Clason and Kunisch. They introduce the terms *multi-bang control* and *generalized multi-bang principle* for the control variable $\boldsymbol{v}$ (u in their notation) if $\boldsymbol{v}(x) \in \{v_1, \ldots, v_M\}$ for a.a. $x \in \bar{\Omega}$. In contrast to [22, (1.1)], the term $\beta \int_\Omega \prod_{i=1}^{M} |\boldsymbol{v} - v_i|_0$ is missing in (MIOCP) and the box constraint $\boldsymbol{v}(x) \in [v_1, v_M]$ is replaced by $\boldsymbol{v}(x) \in \{v_1, \ldots, v_M\}$.



(a) Relaxed controls $\boldsymbol{\alpha}_1$, $\boldsymbol{\alpha}_2$, $\boldsymbol{\alpha}_3$ (left) and realizations v_1, v_2, v_3 (right).



(b) $d^{(n)}(\boldsymbol{\omega}^{(n)}, \boldsymbol{\alpha})$ for (SUR) and bounds for (SUR) and (SUR-VC) from Chapter 6.

Figure 10.7: Relaxed control (left) and behavior of SUR (right) for Example 10.1. The figure has been published as Figure 4 in [79].

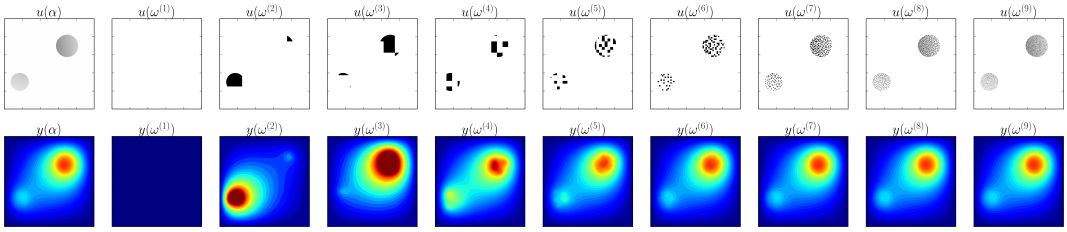


Figure 10.8: Illustration of $\sum_i \boldsymbol{\omega}_i^{(n)} v_i \rightharpoonup^* \sum_i \boldsymbol{\alpha}_i v_i$ (top) and $y(\boldsymbol{\omega}^{(n)}) \to y(\boldsymbol{\alpha})$ (bottom). The figure has been published as Figure 5 in [79].

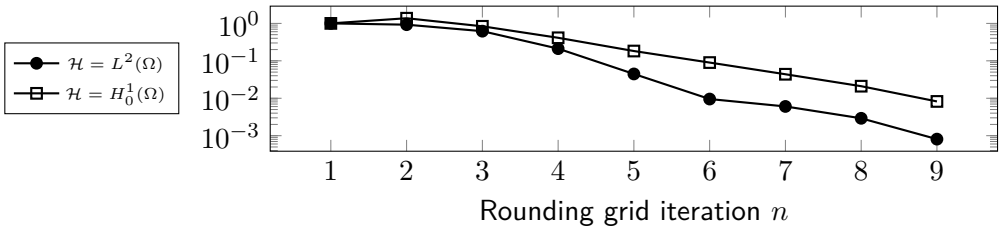


Figure 10.9: Relative state approximation error $\frac{\|y(\boldsymbol{\alpha}) - y(\boldsymbol{\omega}^{(n)})\|_{\mathcal{H}}}{\|y(\boldsymbol{\alpha})\|_{\mathcal{H}}}$.

Therein, we have the setting $|t|_0 \coloneqq 1 - \delta_{t0}$ using the real-valued Kronecker δ and hence, $\beta \int_\Omega \prod_{i=1}^M |\boldsymbol{v} - v_i|_0$ promotes $\{v_1, \dots, v_M\}$-valued controls. Thus, we may interpret (MIOCP) as a limit problem of [22, (1.1)] for $\beta \to \infty$.

Reformulation and relaxation The continuous relaxation of (MIOCP) from *partial outer convexification* reads

$$\begin{aligned} \min_{y,\boldsymbol{v}} \ & \tfrac{1}{2}\|y - y_d\|_{L^2}^2 + \tfrac{\gamma}{2}\left\|\textstyle\sum_{i=1}^M \boldsymbol{\alpha}_i v_i\right\|_{L^2}^2 \\ \text{s.t.} \ & -\Delta_D y = \textstyle\sum_{i=1}^M \boldsymbol{\alpha}_i v_i \\ & \boldsymbol{\alpha}(x) \in \operatorname{conv} \mathbb{S}^M \text{ for a.a. } x \in \Omega. \end{aligned} \tag{P RC1}$$

As in Section 10.2, we reformulate (P RC1) equivalently to

$$\begin{aligned} \min_{y,\boldsymbol{v}} \ & \tfrac{1}{2}\|y - y_d\|_{L^2}^2 + \tfrac{\gamma}{2}\|\boldsymbol{v}\|_{L^2}^2 \\ \text{s.t.} \ & -\Delta_D y = \boldsymbol{v} \\ & \boldsymbol{v}(x) \in [v_1, v_M] \text{ for a.a. } x \in \Omega, \end{aligned} \tag{P RC2}$$

which can be solved more easily and compute $\boldsymbol{\alpha}$ from $\boldsymbol{v}$ after solving (P RC2). The problem (P RC1) is ill-posed because the value of $\boldsymbol{v}$ can in general be represented with more than one convex combination of the v_i and the particular outcome of (SUR) depends on the chosen representation (if we do not have the case $\boldsymbol{v}(x) \in \{v_1, v_M\}$ for a.a. $x \in \Omega$). In particular, a different $\boldsymbol{\alpha}$ is approximated by the $\boldsymbol{\omega}^{(n)}$ in the weak* topology. Again, we represent a value $\boldsymbol{v}(x)$ for $x \in \Omega$ as the convex combination of its two neighboring points in $\{v_1, \dots, v_M\}$.

L^1-regularized problems are known to yield large parts of the domain in which the control is zero. If there exists $v^* \in \{v_1, \dots, v_M\}$ such that $\boldsymbol{v}(x) = v^*$ in large parts of the domain can be assumed a priori, we suggest to solve an L^1-regularized problem with regularizer $\|\boldsymbol{v} - v^*\|_{L^1}$ in the relaxation. We demonstrate this by including the (relaxed) control problem

$$\begin{aligned} \min_{y,\boldsymbol{v}} \ & \tfrac{1}{2}\|y - y_d\|_{L^2}^2 + \tfrac{\gamma}{2}\|\boldsymbol{v}\|_{L^2}^2 + \eta\|\boldsymbol{v}\|_{L^1} \\ \text{s.t.} \ & -\Delta_D y = \boldsymbol{v} \\ & \boldsymbol{v}(x) \in [v_1, v_M] \text{ for a.a. } x \in \Omega \end{aligned} \tag{P RC3}$$

into our experiments. If $\gamma > 0$, the L^2-term improves the regularity of the solution without the need to smooth the L^1-term. Problems of the form (P RC3) are investigated in [109, 117]. We solve the discretization of (P RC3) by means of the *active set method* from [109] with $\gamma \ll \eta$ to stress the effect of the L^1-regularization over the L^2-regularization.

Approximation results As the objective depends on $\boldsymbol{\alpha}$ in (P RC2), the assumptions of Corollary 4.6 are satisfied, but Assumption 4.1 does not hold for the whole objective. For an admissible sequence of refined rounding grids, we have $\boldsymbol{v}^{(n)} \rightharpoonup \boldsymbol{v}$ but the norm $\|\cdot\|_{L^2}$ is only weakly lower semi-continuous, which yields

$$\liminf \frac{\gamma}{2}\|\boldsymbol{v}^{(n)}\|_{L^2} \geq \frac{\gamma}{2}\|\boldsymbol{v}\|_{L^2}^2$$

with equality only if $\boldsymbol{v}^{(n)} \to \boldsymbol{v}$, i.e. the optimal relaxed control being already $\{v_1, \dots, v_M\}$-valued. Thus, we expect norm convergence of the tracking type summand and convergence of the L^2-regularization term to a suboptimal value.

We use the tracking target y_d and the bangs $v_1 = -2, \dots, v_5 = 2$ from [22], which allows to employ their code for plausibility checks. We solve (P RC2), with $\gamma = 10^{-3}$, and (P RC3), with $\gamma = 10^{-5}$ and $\eta = 5 \cdot 10^{-4}$, on a predefined triangular mesh with first order Lagrange finite elements and use piecewise constant discontinuous Galerkin elements for the discrete-valued controls $\boldsymbol{v}^{(n)}$. We start with the rounding grid consisting of four square cells and refine uniformly, see Fig. 10.4. We stop the refinement if the radius of the enclosing circle of a square cell in the rounding grid matched that of two triangles of the finite element mesh. The objective values and the corresponding relative suboptimalities of the iterates are given in Table 10.3 for (P RC2) and Table 10.4 for (P RC3).

Table 10.3: Convergence of objective, relative objective difference and objective summands for (P RC2) with $\gamma = 10^{-3}$. The table has been published as Table 2 in [79].

Iteration	Obj. value	Rel. obj. difference	$\frac{1}{2}\|y^{(n)} - y_d\|_{L^2}^2$	$\frac{\gamma}{2}\|\boldsymbol{v}^{(n)}\|_{L^2}^2$
1	1.7274×10^{-2}	8.40×10^{-2}	1.652×10^{-2}	7.500×10^{-4}
2	1.6913×10^{-2}	6.13×10^{-2}	1.591×10^{-2}	1.000×10^{-3}
3	1.6261×10^{-2}	2.04×10^{-2}	1.534×10^{-2}	9.219×10^{-4}
4	1.6054×10^{-2}	7.41×10^{-3}	1.513×10^{-2}	9.258×10^{-4}
5	1.6022×10^{-2}	5.37×10^{-3}	1.510×10^{-2}	9.214×10^{-4}
6	1.5996×10^{-2}	3.75×10^{-3}	1.508×10^{-2}	9.203×10^{-4}
7	1.5991×10^{-2}	3.48×10^{-3}	1.507×10^{-2}	9.200×10^{-4}
8	1.5990×10^{-2}	3.37×10^{-3}	1.507×10^{-2}	9.200×10^{-4}
9	1.5989×10^{-2}	3.35×10^{-3}	1.507×10^{-2}	9.200×10^{-4}
Relaxed	1.5936×10^{-2}	0	1.507×10^{-2}	8.668×10^{-4}

As expected, the tracking type summand of the objective converges to the value of the relaxed problem in both cases while the regularizer converges to a suboptimal value for (P RC2). For (P RC3), this happens as well, but the suboptimality is smaller because the $\boldsymbol{v}^{(n)}$ approximate $\boldsymbol{v}$ closely in the norm-topology. We visualize the relaxed control $\boldsymbol{v}$ and the discrete-valued approximants $\boldsymbol{v}^{(n)}$, which are reconstructed from the outputs of (SUR) in Fig. 10.10 for (P RC2) and in Fig. 10.11 for (P RC3). The closer approximation in the norm topology for (P RC3) can be perceived visually.

Table 10.4: Convergence of objective, relative objective difference and objective summands for (P RC3) with $\gamma = 10^{-5}$ and $\eta = 5 \cdot 10^{-4}$. The table has been published as Table 3 in [79].

It.	Obj. value	Rel. obj. difference	$\frac{1}{2}\|y^{(n)} - y_d\|_{L^2}^2$	$\frac{\gamma}{2}\|\boldsymbol{v}^{(n)}\|_{L^2}^2 + \eta\|\boldsymbol{v}^{(n)}\|_{L^1}$
1	1.6467×10^{-2}	6.30×10^{-2}	1.570×10^{-2}	7.6500×10^{-4}
2	1.6273×10^{-2}	5.05×10^{-2}	1.564×10^{-2}	6.3750×10^{-4}
3	1.5709×10^{-2}	1.41×10^{-2}	1.507×10^{-2}	6.3750×10^{-4}
4	1.5564×10^{-2}	4.78×10^{-3}	1.493×10^{-2}	6.3348×10^{-4}
5	1.5506×10^{-2}	1.03×10^{-3}	1.487×10^{-2}	6.3499×10^{-4}
6	1.5497×10^{-2}	4.26×10^{-4}	1.486×10^{-2}	6.3462×10^{-4}
7	1.5493×10^{-2}	1.46×10^{-4}	1.486×10^{-2}	6.3458×10^{-4}
8	1.5491×10^{-2}	3.02×10^{-5}	1.486×10^{-2}	6.3459×10^{-4}
9	1.5491×10^{-2}	7.18×10^{-6}	1.486×10^{-2}	6.3460×10^{-4}
Rel.	1.5490×10^{-2}	0	1.486×10^{-2}	6.3456×10^{-4}

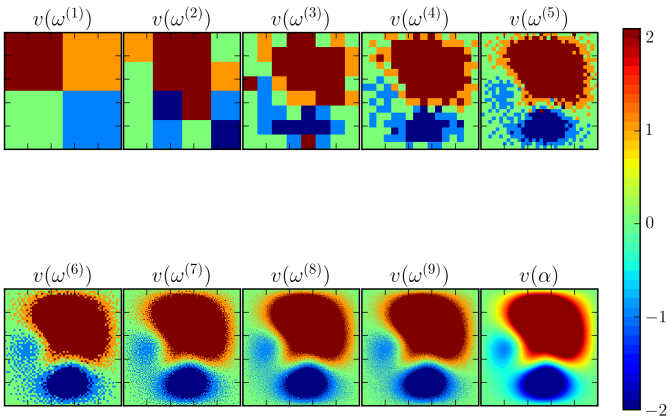


Figure 10.10: Visualization of the weak convergence $\boldsymbol{v}^{(n)} \rightharpoonup \boldsymbol{v}$ for $\boldsymbol{v}$ solving (P RC2). The figure has been published as Figure 6 in [79].

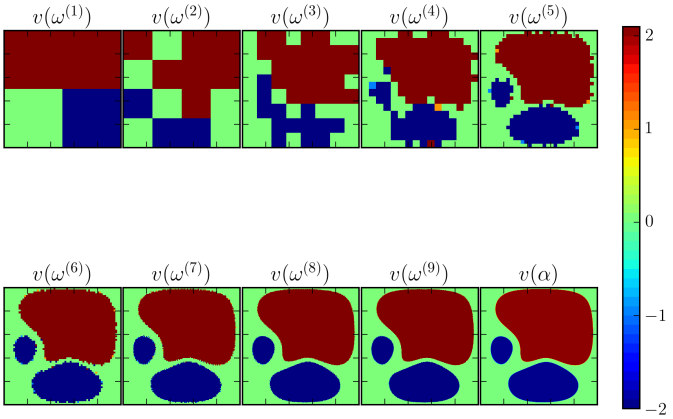


Figure 10.11: Visualization of the weak convergence $\boldsymbol{v}^{(n)} \rightharpoonup \boldsymbol{v}$ for $\boldsymbol{v}$ solving (P RC3). The figure has been published as Figure 7 in [79].

10.4 State vector approximation for the fractional Laplacian

Let $d \geq 2$, $\Omega \subset \mathbb{R}^d$ be a bounded domain with Lipschitz boundary $\partial\Omega$. We validate (1.1) for the elliptic state equation

$$(-\Delta)^s y = f, \quad y|_{\partial\Omega} = 0 \tag{10.2}$$

with $f \in L^2(\Omega)$ and $s \in (0,1)$. It remains to analyze (c) as the rest follows as in Section 10.3. From now on, the boundary condition is considered as part the operator $(-\Delta_D)^s$ to which we refer as *fractional Laplacian* as in [86]. We diagonalize the Dirichlet Laplacian by means of its Fourier series with orthonormal basis functions $(e_n)_n$, i.e.

$$(-\Delta_D)y = \sum_{n=1}^{\infty} \lambda_n \langle y, e_n\rangle_{L^2} e_n.$$

This gives rise to the following definition of the fractional Laplacian

$$(-\Delta_D)^s y \coloneqq \sum_{n=1}^{\infty} \lambda_n^s \langle y, e_n\rangle_{L^2} e_n \tag{10.3}$$

for all $y \in L^2(\Omega)$ such that expression constitutes an L^2-function. Note that for $-\Delta_D$, we have a positive spectrum $(\lambda_n)_n \subset \mathbb{R}_+$. Solutions of (10.2) can be discussed using fractional Sobolev spaces, see e.g. [2, 29, 76], which we briefly introduce following [76].

Let $\mathcal{H}$ and $\mathcal{K}$ be Hilbert spaces with inner products $\langle\cdot,\cdot\rangle_{\mathcal{H}}$ and $\langle\cdot,\cdot\rangle_{\mathcal{K}}$ and a continuous and dense embedding $\mathcal{H} \hookrightarrow \mathcal{K}$. One can construct a self-adjoint, positive and unbounded operator $\Lambda : \mathcal{H} = D(\Lambda) \to \mathcal{K}$ such that

$$\langle u, v\rangle_{\mathcal{H}} = \langle \Lambda u, \Lambda v\rangle_{\mathcal{K}}$$

for all $u, v \in \mathcal{H}$. We define interpolation spaces by means of the domain of Λ.

Definition 10.2 (Def. 2.1 in [76]). *Let Λ be constructed as above and $s \in [0,1]$. Then, we define the interpolation space $[\mathcal{H},\mathcal{K}]_s \coloneqq D(\Lambda^{1-s})$.*

Furthermore, we introduce the following space and its norm

$$\mathbb{H}^s(\Omega) \coloneqq \left\{ w \in L^2(\Omega) : \sum_{n=1}^{\infty} \lambda_n^s |\langle w, e_n\rangle_{L^2}|^2 < \infty \right\},$$

$$\|w\|_{\mathbb{H}^s(\Omega)} \coloneqq \left(\|w\|^2_{L^2(\Omega)} + \sum_{n=1}^{\infty} \lambda_n^s |\langle w, e_n\rangle_{L^2}|^2 \right)^{\frac{1}{2}},$$

which enable us to conclude well-definedness of the variational form of $(-\Delta_D)^s$, which is essentially the Cauchy–Schwarz inequality for $\mathbb{H}^s(\Omega)$.

Lemma 10.3. *Let $y \in \mathbb{H}^s(\Omega)$, $v \in \mathbb{H}^s(\Omega)$. Then, the variational form of $(-\Delta_D)^s$,*

$$\langle (-\Delta_D)^s y, v\rangle_{L^2} = \sum_{n=1}^{\infty} \lambda_n^s \langle y, e_n\rangle_{L^2} \langle e_n, v\rangle_{L^2},$$

is well-defined, i.e. continuous in v and y.

Proof.

$$|\langle(-\Delta_D)^s y, v\rangle_{L^2}|^2 = \left|\sum_{n=1}^{\infty} \lambda_n^s \langle y, e_n\rangle_{L^2}\langle e_n, v\rangle_{L^2}\right|^2 \leq \left(\sum_{n=1}^{\infty} \lambda_n^{\frac{s}{2}} |\langle y, e_n\rangle_{L^2}| \lambda_n^{\frac{s}{2}} |\langle v, e_n\rangle_{L^2}|\right)^2$$
$$\underset{\text{C.S.}}{\leq} \left(\sum_{n=1}^{\infty} \lambda_n^s |\langle y, e_n\rangle_{L^2}|^2\right)\left(\sum_{n=1}^{\infty} \lambda_n^s |\langle v, e_n\rangle_{L^2}|^2\right) = \|y\|_{\mathbb{H}^s}^2 \|v\|_{\mathbb{H}^s}^2$$

□

Moreover, $\mathbb{H}^s(\Omega)$ is the domain of $(-\Delta_D)^{\frac{s}{2}} : D((-\Delta_D)^{\frac{s}{2}}) \to L^2(\Omega)$.

Proposition 10.4. $\mathbb{H}^s(\Omega) = \left\{w \in L^2(\Omega) : \left\|(-\Delta_D)^{\frac{s}{2}} w\right\|_{L^2} < \infty\right\}$

Proof. Let $m_1 < m_2$ and consider, cf. (10.3),

$$\sum_{n=1}^{m_2} \lambda^{\frac{s}{2}} \langle y, e_n\rangle_{L^2} e_n - \sum_{n=1}^{m_1} \lambda^{\frac{s}{2}} \langle y, e_n\rangle_{L^2} e_n = \sum_{n=m_1+1}^{m_2} \lambda^{\frac{s}{2}} \langle y, e_n\rangle_{L^2} e_n$$

from which we deduce the identity

$$\left\|\sum_{n=m_1+1}^{m_2} \lambda^{\frac{s}{2}} \langle y, e_n\rangle_{L^2} e_n\right\|_{L^2}^2 = \left\langle \sum_{n=m_1+1}^{m_2} \lambda_n^{\frac{s}{2}} \langle y, e_n\rangle_{L^2} e_n, \sum_{n=m_1+1}^{m_2} \lambda_n^{\frac{s}{2}} \langle y, e_n\rangle_{L^2} e_n \right\rangle_{L^2}$$
$$= \sum_{n=m_1+1}^{m_2} \lambda_n^s \langle y, e_n\rangle_{L^2}^2.$$

Consequently, the sequence $\left(\sum_{n=1}^{m} \lambda^{\frac{s}{2}} \langle y, e_n\rangle_{L^2} e_n\right)_m$ is Cauchy in $L^2(\Omega)$ if and only if $\left(\sum_{n=1}^{m} \lambda_n^s \langle y, e_n\rangle_{L^2}^2\right)_m$ is Cauchy in $\mathbb{R}$, which proves the claim. □

We denote the topological dual of $\mathbb{H}^s(\Omega)$ by $\mathbb{H}^{-s}(\Omega)$. Before we are able to introduce weak solutions of (10.2), we have to make a note on the norm of $\mathbb{H}^s(\Omega)$.

Lemma 10.5. *There exists $C > 0$ such that for every $w \in \mathbb{H}^s(\Omega)$*

$$\left(\sum_{n=1}^{\infty} |\langle w, e_n\rangle_{L^2}|^2\right)^{\frac{1}{2}} = \|w\|_{L^2} \leq C \left(\sum_{n=1}^{\infty} \lambda_n^s |\langle w, e_n\rangle_{L^2}|^2\right)^{\frac{1}{2}}.$$

Proof. The operator $(-\Delta_D)^{-1} : L^2(\Omega) \to H_0^1(\Omega)$ is compact. The Hilbert-Schmidt theorem (see Theorem B.5) gives that its spectrum is countable and discrete with 0 being the only accumulation point. Furthermore, the $-\Delta_D$ is self-adjoint and positive, see e.g. [103, Sect. 10.6.1]. Noticing $(-\Delta_D)(-\Delta_D)^{-1} y = y$ for $y \in L^2(\Omega)$ and $(-\Delta_D)^{-1}(-\Delta_D) y = y$ for $y \in H_0^1(\Omega)$, all eigenfunctions of $-\Delta_D$ are eigenfunctions of $(-\Delta_D)^{-1}$ with reciprocal spectral values and the spectral values $(\lambda_n)_n$ of $(-\Delta_D)$ are positive and satisfy $0 < \lambda_1 < \lambda_2 < \ldots$ with $\lambda_n \to \infty$. Thus, there exists $n_0 \in \mathbb{N}$ such that $\lambda_n \geq 1$ for $n \geq n_0$ giving

$$\sum_{n=n_0}^{\infty} |\langle w, e_n\rangle_{L^2}|^2 \leq \sum_{n=n_0}^{\infty} \lambda_n^s |\langle w, e_n\rangle_{L^2}|^2.$$

Setting $c_1 \coloneqq \frac{1}{\min\{\lambda_1^s,\dots,\lambda_{n_0-1}^s\}} > 0$, we obtain

$$\sum_{n=0}^{n_0-1} |\langle w, e_n\rangle_{L^2}|^2 \leq c_1 \sum_{n=0}^{n_0-1} \lambda_n^s |\langle w, e_n\rangle_{L^2}|^2.$$

Using $C \coloneqq (\max\{1, c_1\})^{\frac{1}{2}}$, the desired estimate holds. □

We state the *standard* result for the homogeneous Dirichlet problem (10.2).

Proposition 10.6 (Solution of (10.2))**.** *Let $f \in \mathbb{H}^{-s}(\Omega)$. Then, there exists a unique $y \in \mathbb{H}^s(\Omega)$ such that*

$$\langle (-\Delta_D)^{\frac{s}{2}} y, (-\Delta_D)^{\frac{s}{2}} v\rangle_{L^2} = \langle f, v\rangle_{\mathbb{H}^{-s},\mathbb{H}^s} \text{ for all } v \in \mathbb{H}^s(\Omega).$$

Proof. The claim follows from Lax–Milgram lemma (see Theorem B.6) if

$$(v, y) \mapsto \langle (-\Delta_D)^{\frac{s}{2}} v, (-\Delta_D)^{\frac{s}{2}} y\rangle_{L^2}$$

constitutes a continuous bilinear form that is coercive on $\mathbb{H}^s(\Omega)$, i.e.

$$\langle (-\Delta_D)^{\frac{s}{2}} v, (-\Delta_D)^{\frac{s}{2}} v\rangle_{L^2} \geq c\|v\|_{\mathbb{H}^s}^2$$

for some $c > 0$ and all $v \in \mathbb{H}^s(\Omega)$. Continuity follows from (the proof of) Lemma 10.3 and bilinearity is straightforward. Lemma 10.5 establishes the coercivity. □

Proposition 10.7 (Compact embedding)**.** *Let $s \in (0, 1)$. Then, $\mathbb{H}^s(\Omega) \hookrightarrow^c L^2(\Omega)$.*

Proof. Definition 10.2 and Proposition 10.4 give $\mathbb{H}^s(\Omega) = [H_0^1(\Omega), L^2(\Omega)]_{1-s}$. Combining these characterizations with the compact embeddings for interpolation spaces [76, Thm 16.2] yields the claim. □

Theorem 10.8. *Let $\boldsymbol{\alpha}$ be a relaxed control. Let $f_i \in L^2(\Omega)$ for $i \in \{1, \dots, M\}$. Let $y \in \mathbb{H}^s(\Omega)$ be a weak solution of (10.2) with $f = \sum_{i=1}^M \boldsymbol{\alpha}_i f_i$. Let $(\boldsymbol{\omega}^{(n)})_n \subset L^\infty(\Omega, \mathbb{R}^M)$ satisfy $\boldsymbol{\omega}^{(n)} \overset{*}{\rightharpoonup} \boldsymbol{\alpha}$ and $(y^{(n)})_n \subset \mathbb{H}^s(\Omega)$ be the corresponding weak solutions of (10.2) with $f = \sum_{i=1}^M \boldsymbol{\omega}_i^{(n)} f_i$ for $n \in \mathbb{N}$. Then, $y^{(n)} \to y$ in $\mathbb{H}^s(\Omega)$.*

Proof. The well-definedness, i.e. the existence of the weak solutions $y(\boldsymbol{\alpha}) \in \mathbb{H}^s(\Omega)$ and $y(\boldsymbol{\omega}^{(n)}) \in \mathbb{H}^s(\Omega)$, follows by virtue of Proposition 10.6. From the prerequisites and Proposition 10.7, we obtain $\sum_{i=1}^M \boldsymbol{\omega}_i^{(n)} f_i \to \sum_{i=1}^M \boldsymbol{\alpha}_i f_i$ in $\mathbb{H}^{-s}(\Omega)$. Finally, the Lax–Milgram lemma (Theorem B.6) gives continuity of the solution of (10.2). □

Remark 10.9. *For further reading on different characterizations of $\mathbb{H}^s(\Omega)$, the interpolation spaces $[H_0^1(\Omega), L^2(\Omega)]_{1-s}$ and $[H^1(\Omega), L^2(\Omega)]_{1-s}$ and their norms, we refer to* [76, Chap. 1.9, 1.11] *and* [29]. *In* [76]*, smooth boundaries are assumed. However, Lipschitz boundary conditions also comply with the involved spaces, see* [115, p. 164] *for the argument based on Stein's extension theorem* [110, Sect. VI.§3.1 Thm 5].

We demonstrate (1.1) numerically using a control from Section 10.3, which is again approximated weakly using (SUR) on successively refined grids and along orderings induced by successive Hilbert curve iterates as in Section 10.3. To solve (10.3) for $s \in (0,1)$, we employ the quadrature rules from [4], which are based on the detailed considerations in [15]. The convergence of control and state vectors is visualized in Figs. 10.12 and 10.13 for the choices $s = 0.4$ and $s = 0.7$.

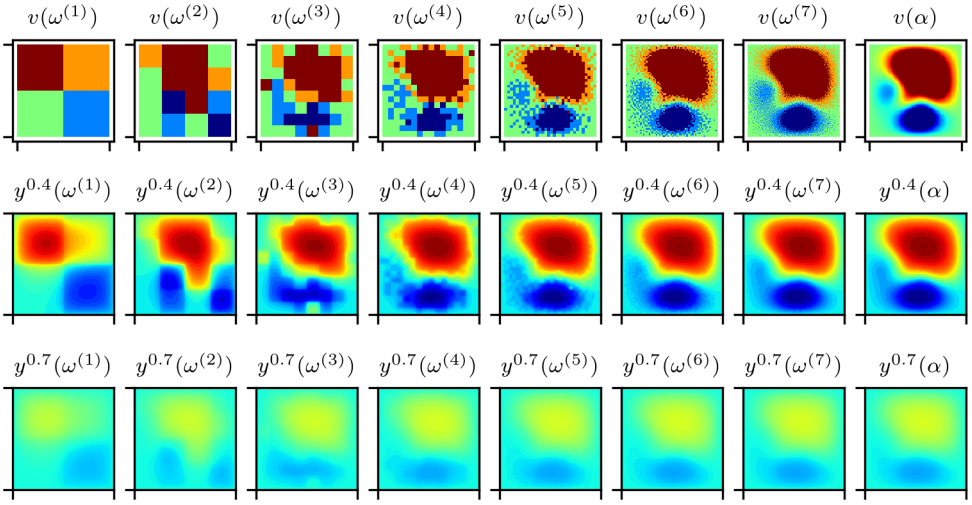


Figure 10.12: $\boldsymbol{v}(\boldsymbol{\omega}^{(n)}) \rightharpoonup \boldsymbol{v}(\boldsymbol{\alpha})$ (top) and $y(\boldsymbol{\omega}^{(n)}) \to y(\boldsymbol{\alpha})$ (center, bottom) for $s = 0.4$, 0.7 with $y(\boldsymbol{\omega}^{(n)})$ solving (10.3) for $f = \boldsymbol{v}(\boldsymbol{\omega}^{(n)})$ and $y(\boldsymbol{\alpha})$ for $f = \boldsymbol{v}(\boldsymbol{\alpha})$.

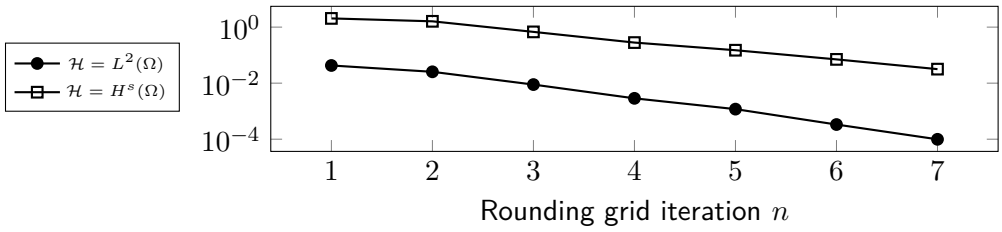


Figure 10.13: Relative state approximation error $\frac{\|y(\boldsymbol{\alpha}) - y(\boldsymbol{\omega}^{(n)})\|_{\mathcal{H}}}{\|y(\boldsymbol{\alpha})\|_{\mathcal{H}}}$ for $(-\Delta_D)^{0.7}$.

Chapter 11

Algorithmic framework

11.1 Approximation algorithm

Algorithm 11.1 (MIOCP) Approximation

Require: Let Assumption 4.1 hold.

Require: Let $\mathcal{Y} \ni y \mapsto c(y, u, v_i) \in L^\infty(\Omega_T)$ be continuous for all $u \in \mathcal{U}$ and $i \in \{1, \dots, M\}$.

Input: Approximation tolerance $\delta_{\max} > 0$.

Input: Initial grid $S^{(0)}$

$y, u, \boldsymbol{\alpha} \leftarrow$ SOLVE((RC))
$n \leftarrow 0$
do
  if $n > 0$ **then**
    $S^{(n)} \leftarrow$ REFINE$(S^{(n-1)})$
  end if
  $\boldsymbol{\omega}^{(n)} \leftarrow$ (SUR-VC)$(\boldsymbol{\alpha}, S^{(n)})$
  $y^{(n)} \leftarrow S_R(u, \boldsymbol{\omega}^{(n)})$
  $\delta_1 \leftarrow J(y^{(n)}, u) - J(y, u)$
  $\delta_2 \leftarrow \|\min\{0, \min_i \boldsymbol{\omega}_i c(y, u, v_i)\}\|_{L^\infty}$
  $n \to n + 1$
while $\delta_1 \geq \delta_{\max}$ or $\delta_2 \geq \delta_{\max}$

We introduce Algorithm 11.1, which synthesizes our findings into an algorithm. In the previous chapters, we have shown (among other things) that our chain of approximation arguments (1.1) holds constructively if the algorithm (SUR-VC) is used for rounding on an admissible sequence of refined rounding grids provided suitable regularity assumptions on the state equation and the pointwise a.e. defined mixed constraint hold for (MIOCP). This gives termination of Algorithm 11.1 within finitely many steps for a strictly positive tolerance $\delta_{\max} > 0$, which we summarize in the following theorem.

Theorem 11.1. *Let the prerequisites of Algorithm 11.1 hold and $(S^{(n)})_n$ be an admissible sequence of refined rounding grids and $(y, u, \boldsymbol{\alpha}) \in \mathcal{F}_{(\mathsf{RC})}$ be optimal for* (RC)*. Then, Algorithm 11.1 terminates after finitely many iterations with $(y^{(n)}, u, \boldsymbol{\omega}^{(n)}) \in \mathcal{F}_{(\mathsf{BC}_{\delta_{\max}})}$ satisfying*

$$J(y^{(n)}, u) - J(y, u) < \delta_{\max}.$$

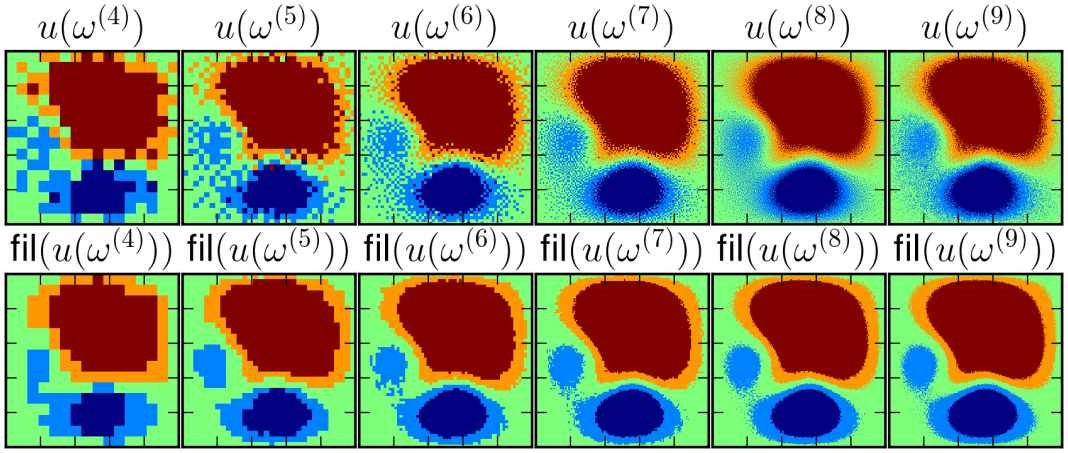


Figure 11.1: Reconstructed controls for (SUR) along successive Hilbert curve iterates (top) and after the application of a median filter as post-processing step (bottom).

11.2 Postprocessing

As seen in Chapter 10, the (SUR-GEN) algorithms may produce chattering. In particular if the mixed constraint is absent and (SUR) is used, it might make sense to extend Algorithm 11.1 by heuristic post-processing steps to reduce the chattering and obtain more implementation-friendly controls. The heuristic step can be accepted or rejected depending on whether the objective value and / or the infeasibility were improved.

The median filter can serve for this purpose[8] if a meaningful ordering of $v_1, \ldots, v_M$ exists as in the setting of Section 10.3.3. For this setting, median filtering improves the objective value for all refinements, see Table 11.1. The filter has been implemented as the median of the 9 cells in a 3×3 window centered at the considered cell. Figure 11.1

Table 11.1: Objective and relative suboptimality with and without median filtering.

	Before postprocessing		After postprocessing	
Iteration	Objective	Rel. suboptim.	Objective	Rel. suboptim.
1	1.7274×10^{-2}	8.40×10^{-2}	1.6524×10^{-2}	3.69×10^{-2}
2	1.6913×10^{-2}	6.13×10^{-2}	1.6225×10^{-2}	1.81×10^{-2}
3	1.6261×10^{-2}	2.04×10^{-2}	1.6003×10^{-2}	4.20×10^{-3}
4	1.6054×10^{-2}	7.41×10^{-3}	1.5945×10^{-2}	5.46×10^{-4}
5	1.6022×10^{-2}	5.37×10^{-3}	1.5987×10^{-2}	3.22×10^{-3}
6	1.5996×10^{-2}	3.75×10^{-3}	1.5975×10^{-2}	2.46×10^{-3}
7	1.5991×10^{-2}	3.48×10^{-3}	1.5971×10^{-2}	2.19×10^{-3}
8	1.5990×10^{-2}	3.37×10^{-3}	1.5973×10^{-2}	2.33×10^{-3}
9	1.5989×10^{-2}	3.35×10^{-3}	1.5973×10^{-2}	2.32×10^{-3}
Relaxed	1.5936×10^{-2}	0	-	-

visualizes the iterates before and after applying the median filter.

[8] We acknowledge that the idea of employing the median filter is due to Dirk Lorenz, TU Braunschweig.

Chapter 12

Discussion

We have analyzed and evaluated a chain of approximation arguments in the previous chapters to obtain approximation properties w.r.t. driving the mesh sizes of a sequence of rounding grids to zero. Our arguments generalize to different rounding algorithms by checking the following two questions:

1. Does the integrality gap tend to zero if the mesh size tends to zero?
2. Is the control-to-state operator of the state equation completely continuous?

Situations may occur, in which the control-to-state operator has weaker regularity. In case of weak-weak continuity, $\boldsymbol{\omega}^{(h)} \rightharpoonup \boldsymbol{\alpha}$ implies $y(\boldsymbol{\omega}^{(h)}) \rightharpoonup y(\boldsymbol{\alpha})$ and we will most often obtain a gap between the objective value of (RC) and the limit of the objective for the roundings because objectives of OCPs are usually weakly lower semi-continuous but not weakly continuous (weak-norm continuous).

12.1 Summary of the approximation arguments

We have investigated the relationships between (MIOCP), (BC$_\delta$) and (RC) as visualized in Fig. 1.1. The results of our investigation are summarized in Fig. 12.1.

The statements in Chapter 4 generalize and unify the results from [100, Cor. 6& 8], [71, Thm 6.7], [68, Thm 3.6], [51, Thm 1], [80, Prop. 2.5] and [79, Thm 5.1, Cor. 5.2] because these works all show that the considered problem classes satisfy Assumption 4.1, albeit not explicitly and in different ways. The computational results in Chapter 10 strengthen our claim that the proposed methodology provides a computationally efficient way to compute discrete-valued functions without the need to use discrete optimization algorithms which might have problems with the high number of variables when fine discretizations of the state equation and in particular the control variables are desired. In particular, we achieve a constructive way to compute a minimizing sequence to the optimum. However, we highlight a shortcoming in the presented theory: to compute solutions of (RC) numerically, it is often necessary to introduce regularization as

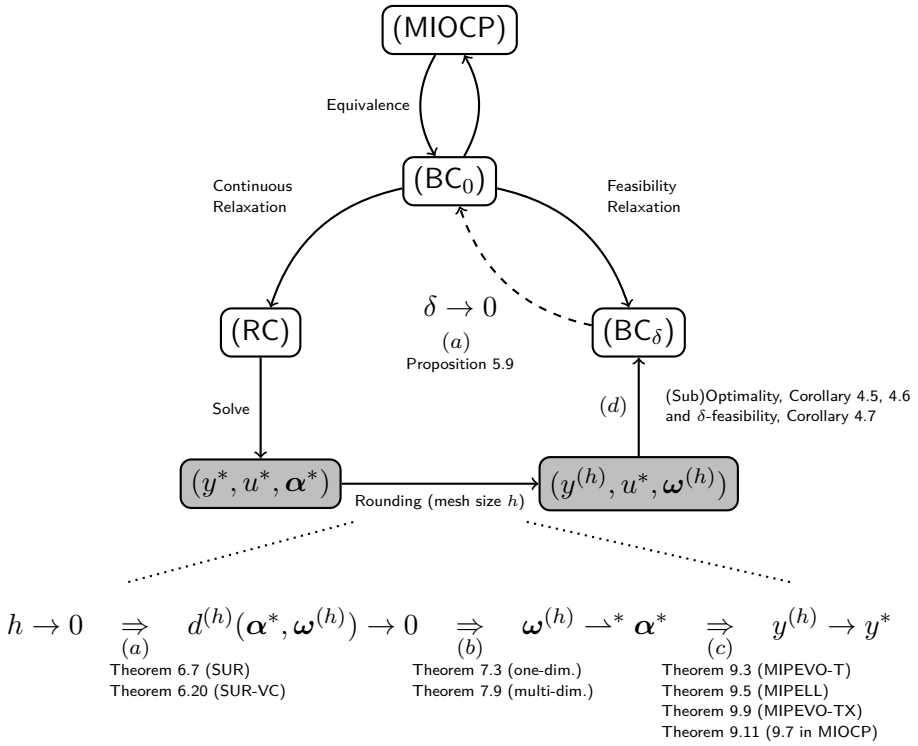


Figure 12.1: Investigated relationship between (MIOCP), (BC$_\delta$) and (RC).

the problems are usually not strictly convex. Common regularizers like powers of L^p-norms are not weakly continuous, but only weakly lower semi-continuous, which yields a bounded suboptimality, which can be controlled by a coefficient, but which may also be fixed a priori.

For the sake of honesty, we highlight that the analysis in [50, 51, 68, 71, 100] gives an a priori estimate, which we lose after removing differentiability assumptions in Section 9.1.

12.2 Optimal binary controls

One may ask what happens if the solution of (RC) already solves (BC$_0$) and when this can be expected. Deckelnick and Hinze make an observation on unregularized elliptic control problems with box-constrained control inputs in [28]. In [28, Sect. 3], they give sufficient conditions on the adjoint state for the solution of the relaxed problem (RC$^{\mathsf{ELL}}$) without control regularization to coincide with the solution of the discrete control problem (BC$_\delta^{\mathsf{ELL}}$) in the case $M = 2$. In particular, the control takes the boundary values of the box everywhere except for a set of measure zero. In this case, they are able to prove error estimates of the form

$$\|y - y_h\|_{L^2} + \|p - p_h\|_{L^\infty} \leq Ch^2.$$

Here, y and p denote the optimal state and adjoint vectors for (P RC1). Their estimate holds in the absence of the regularizing term and y_h and p_h denote the optimal state and adjoint vectors for a discretization with piecewise affine globally continuous finite element functions in the state and adjoint space without explicit discretization of the control space respectively. See [55] for details on this so-called *variational discretization concept*.

12.3 Improving applicability

Considering the improvement from Figure 10.12 to Figure 11.1 with respect to implementation in real machines, we suggest to investigate methods, e.g. using BV-regularization to obtain less fuzzy edges, to obtain binary controls that can be used as inputs in real systems more easily. Furthermore, the proofs for (c) demand that the mesh size is driven to zero. However, for practical purposes, we suggest to consider adaptively refined, in particular non-uniformly refined, rounding grids to avoid fine discretizations and computational overhead for the solution of the state equation in parts of the domain, where this is unnecessary.

Appendix A

Sum formulas

Lemma A.1. *Let $k \in \{1, \dots, M\}$. Then,*

$$(M-k)\sum_{i=0}^{k-1}\frac{1}{M-i} = \sum_{i=0}^{k-1} 1 - \sum_{j=0}^{i}\frac{1}{M-j}$$

Proof.

$$\begin{aligned}(M-k)\sum_{i=0}^{k-1}\frac{1}{M-i} &= \frac{M-k}{M} + \frac{M-1-k+1}{M-1} + \ldots + \frac{M-(k-1)-1}{M-(k-1)} \\ &= 1 - \frac{k}{M} + 1 - \frac{k-1}{M-1} + \ldots + 1 - \frac{1}{M-(k-1)}\end{aligned}$$

We split the k fractions into summands of the form $1/(M-j)$ for $j \in \{0, \dots, k-1\}$. The claim follows from regrouping them into k sums as follows

$$\begin{aligned}(M-k)\sum_{i=0}^{k-1}\frac{1}{M-i} = 1 &- \frac{1}{M} \\ &+ 1 - \frac{1}{M} - \frac{1}{M-1} \\ &+ \ldots \\ &+ 1 - \frac{1}{M} - \ldots - \frac{1}{M-(k-1)}.\end{aligned}$$

□

Lemma A.2 (Appendix of [82])**.** *In Lemma 6.43, we have for $j \in \{2, \dots, |J|-2\}$*

$$d_j^{\kappa_{|J|-1}} = \frac{\overline{\Delta} + \sum_{i=1}^{j+1} d_i^{\kappa_0} 2^{i-1}}{2^{j+1}}$$

Proof. To make the argument more accessible, we abbreviate $d_i \coloneqq d_i^{\kappa_{|J|-1}}$ and $c_i \coloneqq d_i^{\kappa_0}$ for $i \in \{1, \dots, |J|-1\}$ as well as $y_i \coloneqq [\mathbf{\Phi}]^{\mathfrak{s}}_{\kappa_{|J|-1}, j_i^{\kappa_{|J|-1}}}$ and $x_i \coloneqq [\mathbf{\Phi}]^{\mathfrak{s}}_{\kappa_0, j_i^{\kappa_0}}$. We proceed

inductively and obtain

$$d_1 = y_1 - y_2 = \frac{\overline{\Delta} + x_1 + x_2}{2} - \frac{\overline{\Delta} + x_1 + x_2 + 2x_3}{4} = \frac{d_1}{4} + \frac{d_2}{2},$$

from the transformation matrix in the base case. For $j \leq |J| - 2$, we observe the identity

$$\sum_{i=1}^{j} d_i = y_1 - y_{j+1}.$$

We plug in the lines for y_1 and y_{j+1} from the transformation matrix and obtain

$$\begin{aligned}\sum_{i=1}^{j} d_i &= \frac{x_1 + x_2 + \overline{\Delta}}{2} - \frac{x_1 + x_2 + 2x_3 + \cdots + 2^j x_{j+2} + \overline{\Delta}}{2^{j+1}} \\ &= \frac{1}{2^{j+1}} \left((2^j - 1)(x_1 + x_2 + \overline{\Delta}) - 2x_3 - \cdots - 2^j x_{j+2} \right) \\ &= \frac{1}{2^{j+1}} \left((2^j - 1)(x_1 + x_2 + \overline{\Delta}) - 2x_3 - \cdots - (2^{j-1} + 2^j) x_{j+1} + 2^j c_{j+1} \right).\end{aligned}$$

Here, the last equality followed from the definition of c_{j+1}. Plugging in the definitions of c_3,...,c_j and adding the necessary factors in front of the corresponding x_3,... accordingly gives

$$\sum_{i=1}^{j} d_i = \frac{1}{2^{j+1}} \left((2^j - 1)(x_1 + x_2 + \overline{\Delta}) - (2^1 + \ldots + 2^j) x_3 + \sum_{i=3}^{j+1} \left(\sum_{\ell=i-1}^{j} 2^\ell \right) c_i \right)$$

We use the formula $2^j - 1 = 1 + \cdots + 2^{j-1}$ to obtain

$$\begin{aligned}\sum_{i=1}^{j} d_i &= \frac{1}{2^{j+1}} \left((2^j - 1)(x_1 + \overline{\Delta}) + x_2 - 2^j x_2 + \sum_{i=2}^{j+1} \left(\sum_{\ell=i-1}^{j} 2^\ell \right) c_i \right) \\ &= \frac{1}{2^{j+1}} \left(\sum_{\ell=0}^{j-1} 2^\ell \overline{\Delta} + \sum_{i=1}^{j+1} \left(\sum_{\ell=i-1}^{j} 2^\ell \right) c_i \right).\end{aligned}$$

The induction hypothesis gives

$$d_i = \frac{1}{2^{i+1}} \left(\overline{\Delta} + \sum_{\ell=1}^{i+1} 2^{\ell-1} c_\ell \right)$$

for $i \in \{1, \ldots, j-1\}$. Thus, summing from one to $j-1$ and factoring $\frac{1}{2^{j+1}}$ gives

$$\begin{aligned}\sum_{i=1}^{j-1} d_i &= \frac{1}{2^{j+1}} \left(\sum_{i=1}^{j-1} 2^{j-i} \overline{\Delta} + \sum_{i=1}^{j-1} \sum_{\ell=1}^{i+1} 2^{j-i+\ell-1} c_\ell \right) \\ &= \frac{1}{2^{j+1}} \left(\sum_{\ell=1}^{j-1} 2^\ell \overline{\Delta} + \sum_{i=1}^{j} c_i \sum_{k=i}^{j} 2^k \right).\end{aligned}$$

Now the subtraction $d_j = \sum_{i=1}^{j} d_i - \sum_{i=1}^{j-1} d_i$ and a close inspection of the two derived sum formulas yield the claim. □

Appendix B

Utilities from Analysis

B.1 Operators

Definition B.1 (Operator and domain, see [94, Def. 8.1]). *Let X and Y be Banach spaces. We call a pair $(A, D(A))$ a* linear operator *from X to Y if*

1. *$D(A) \subset X$ ($D(A)$ is the **domain** of A),*
2. *$A : D(A) \to Y$ is linear.*

Remark. *To improve readability, we relax this notation and often write A and implicitly assume $(A, D(A))$.*

Definition B.2 (Closed linear operator). *Let X, Y be Banach spaces, $(A, D(A))$ be a linear operator with domain $D(A) \subset X$ and codomain Y. We define $(A, D(A))$ to be **closed** if its graph is closed which means that every sequence $(x_n)_n \subset D(A)$ with $x_n \to x \in X$ and $Ax_n \to y \in Y$ satisfies $x \in D(A)$ and $Ax = y$.*

Definition B.3 (Maximal parabolic regularity). *Let $1 < p < \infty$. Let X be a Banach space. Then, a closed operator $(A, D(A))$ with $D(A) \subset X$ is of maximal $L^p((0,T), X)$-regularity if for all $f \in L^p((0,T), X)$, there exists a unique solution $y \in \mathcal{W}^{1,p}((0,T), X) \cap L^p((0,T), D(A))$ of*

$$\partial_t y + Ay = f, \quad y(0) = 0$$

with $\mathcal{W}^{1,p}((0,T), X) := \{y \in L^p((0,T), X), \partial_t y \in L^p((0,T), X)\}$, where $D(A)$ is equipped with the graph norm.

Definition B.4 (Compact operator). *Let X, Y be Banach spaces. A bounded linear operator $A \in \mathcal{L}(X, Y)$ is called **compact** if for all $B \subset X$ with $\sup_{x \in B} \|x\|_X < \infty$, we have $A(B) \subset\subset Y$.*

Theorem B.5 (Hilbert-Schmidt theorem, Thm 3.2.1 in [107], Thm 8.94 in [94]). *Let $\mathcal{H}$ be a Hilbert space, $A \in \mathcal{L}(\mathcal{H}, \mathcal{H})$ be a positive and compact operator. Then, there exists*

an orthonormal basis $(\varphi_n)_n$ of $\mathcal{H}$ consisting of eigenvectors with associated sequence of eigenvalues $(\lambda_n)_n$ and $N \in \mathbb{N} \cup \{\infty\}$ such that

$$\begin{aligned} A\phi_n &= \lambda_n \phi_n \text{ for all } n \\ \lambda_n &\geq \lambda_{n+1} \text{ for all } n \\ \lambda_n &\to 0 \\ A\phi_m &= 0 \text{ for all } m \geq N \text{ if } N < \infty. \end{aligned}$$

Theorem B.6 (Lax–Milgram lemma, Thm 9.14 in [94], Thm 6.2-1 in [20])**.** *Let $\mathcal{H}$ be a Hilbert space, $B : \mathcal{H} \times \mathcal{H} \to \mathbb{R}$ be a coercive bilinear form satisfying a Cauchy–Schwarz-type inequality, i.e.*

$$|B(v, y)| \leq c\|v\|_{\mathcal{H}}\|y\|_{\mathcal{H}}$$

for all v,$y \in \mathcal{H}$ for some $c > 0$. Then, there exists $C > 0$ such that for every $f \in \mathcal{H}^*$, there exists a unique $y \in \mathcal{H}$ such that

$$B(v, y) = \langle v, f\rangle_{\mathcal{H}^*,\mathcal{H}} \text{ for all } v \in \mathcal{H}.$$

Furthermore, we have the following continuity property

$$\|y\|_{\mathcal{H}} \leq \|f\|_{\mathcal{H}^*}.$$

■

B.2 Strongly continuous semigroups of linear operators

Semigroup theory has been an extensive field of study to analyze evolution equations. We briefly summarize the definitions and statements that are necessary to follow our arguments in Sections 3.2 and 9.1 and Appendix B.6. We lean on the monographs by Arendt et al. [6], Engel and Nagel [35], Pazy [87], Renardy and Rogers [94], and Schweizer [104], to which we also refer for further reading on the subject. Semigroups are generalizations of the (matrix) exponential function. In particular, the following definition contains the properties (of the exponential function) we require to hold.

Definition B.7 (Strongly continuous semigroup)**.** *Let X be a Banach space and consider the family of operators $(S(t))_{t\geq 0} \subset \mathcal{L}(X, X)$. Then, we call $(S(t))_{t\geq 0}$ a **strongly continuous semigroup** or **C_0-semigroup** if the following properties hold:*

1. *$S(0) = I$ (Identity at zero)*
2. *$S(t + s) = S(t)S(s)$ for every $t, s \geq 0$ (Semigroup property)*
3. *$\lim_{t\to 0^+} S(t)x = x$ for every $x \in X$ (Strong continuity / continuity in the orbit in $\mathbb{R}_+$)*

The following norm estimate follows immediately from the semigroup property and is known as *exponential growth bound*.

Proposition B.8 (Exponential growth bound, [35, Prop. 5.5], [87, Chap. 1, Thm 2.2], [6, Thm 3.1.7]). *Let $(S(t))_{t\geq 0}$ be a strongly continuous semigroup. Then, there exist constants $\omega \in \mathbb{R}$ and $M \in [1, \infty)$ such that*

$$\|S(t)\|_{op} \leq M \exp(\omega t)$$

for all $t \in \mathbb{R}_0^+$.

The concept to relate a strongly continuous semigroup to the operator for which we want to apply this generalization of the matrix exponential function is its generator, which is introduced below.

Definition B.9 (Generator of a strongly continuous semigroup). *Let X be a Banach space and $(S(t))_{t\geq 0}$ a C_0-semigroup. We define its **generator** $A : D(A) \to X$ as the linear operator*

$$D(A) \ni x \mapsto Ax \coloneqq \lim_{h\to 0^+} \frac{1}{h}(S(h)x - x)$$

with domain

$$D(A) \coloneqq \left\{ x \in X : \lim_{h\to 0^+} \frac{1}{h}(S(h)x - x) \text{ exists} \right\}.$$

Remark B.10. *This means that the domain of the generator consists of the elements in X of which the orbit map is right-differentiable in 0. This suffices to imply differentiability of the whole orbit map, which is the result of the following proposition.*

Proposition B.11 (Differentiability is right differentiability in 0, [35]). *Let $(S(t))_{t\geq 0}$ be a C_0-semigroup and $x \in X$ be fixed. For the orbit map $\xi_x : [0, \infty) \ni t \mapsto S(t)x \in X$, the following are equivalent:*

1. *ξ_x is right differentiable at $t = 0$.*
2. *ξ_x is differentiable on $\mathbb{R}_0^+$.*

■

The following proposition states that dealing with the generator of a C_0-semigroup is much more comfortable than dealing with some arbitrary unbounded operator. In particular, it is densely defined and shows (sequential) continuity in its graph.

Proposition B.12 (Generators are closed and unique, [35, Thm II.1.4]). *Let $(A, D(A))$ be the generator of a C_0-semigroup $(S(t))_{t\geq 0}$ on X. Then, $(A, D(A))$ is closed, densely defined and determines $(S(t))_{t\geq 0}$ uniquely.* ■

Definition B.13 (Abstract Cauchy Problem)**.** *Let A generate a strongly continuous semigroup. Then, we call the following IVP an* ***Abstract Cauchy Problem (ACP)****.*

$$\frac{\mathrm{d}}{\mathrm{d}t}y(t) = Ay(t) + f(t), \quad y(0) = y_0. \tag{ACP}$$

Definition B.14 (Mild solution of (ACP))**.** *Let X be a Banach space, $(S(t))_{t\geq 0}$ be a strongly continuous semigroup generated by $A : D(A) \to X$. We define the* ***mild solution*** *of* (ACP) *for $y_0 \in X$ and $f \in L^1((0,T),X)$ by means of the Variation of Constants Formula (VOC)*

$$y(t) \coloneqq S(t)y_0 + \int_0^t S(t-s)f(s)\,\mathrm{d}s. \tag{VOC}$$

Proposition B.15 (Well-definedness of (VOC), [6, Sect. 3.1])**.** (VOC) *is well-defined under the prerequisites of Definition B.14. In particular, $y \in C([0,\infty),X)$ and y is uniquely defined. If* (VOC) *admits a classical solution $y \in C^1([0,\infty),X)$, it coincides with the mild solution.* ■

Proposition B.16. *Let Y and U be Banach spaces. Let A generate a C_0-semigroup on X. Let $f_i : Y \times U \to Y$ be uniformly Lipschitz continuous in the first argument and jointly continuous in the first and second argument. Furthermore, let $f_i(0,u) \in L^1((0,T),Y)$ for all $u \in L^2((0,T),U)$. Then,* (3.1) *admits a unique mild solution $y \in C([0,T],Y) =: \mathcal{Y}$ for all $u \in L^2((0,T),U)$, $\boldsymbol{\alpha} \in L^\infty((0,T),\mathbb{R}^M)$ and $y_0 \in Y$.*

Proof. Let $(y,u) \in L^1((0,T),Y\times U)$. As in [91, Cor. 2.2], we approximate with simple functions $(y^{(n)},u^{(n)}) = \sum_{i=1}^{k^{(n)}}(\xi_{i,n},\nu_{i,n})\chi_{A_{i,n}}$ defined on measurable sets $A_{i,n} \subset [0,T]$ and obtain the convergence $f(y^{(n)}(t),u^{(n)}(t)) \to f(x(t),u(t))$ for almost all $t \in [0,T]$ from the joint continuity of the f_i. Furthermore,

$$f_i(y^{(n)}(t),u^{(n)}(t)) = \begin{cases} f_i(\xi_{1,n},\nu_{1,n}) & t \in A_{1,n} \\ \cdots & \\ f_i(\xi_{k^{(n)},n},\nu_{k^{(n)},n}) & t \in A_{k_n,n} \end{cases},$$

which gives the measurability of $t \mapsto f_i(y^{(n)}(t),u^{(n)}(t))$ by [6, Cor 1.1.2]. Furthermore,

$$\begin{aligned} &\int_0^T \|f_i(y(t),u(t))\|_X\,\mathrm{d}t \\ &\qquad \leq T\|f_i(0,u)\|_{L^1} + \int_0^T \|f_i(y(t),u(t)) - f(0,u(t))\|_X\,\mathrm{d}t \\ &\qquad \leq T\|f_i(0,u)\|_{L^1} + L_i\|y\|_{L^1}, \end{aligned}$$

where L denotes the Lipschitz constant for the first argument of f_i, which gives the integrability of $f(v,u)$ for all $v \in C([0,T],Y)$ and $u \in L^1((0,T),U)$. Hölder's inequality gives the integrability of $\sum_{i=1}^M \boldsymbol{\alpha}_i f_i(v,u)$. Combining this with the boundedness of $\|S(t)\|_{op}$ for $t \in [0,T]$, we infer that second summand is a continuous mapping in t if y is integrable. Finally, existence and uniqueness follow with a fixed-point argument, see the monographs by Pazy [87, Chap. 6, Thm 1.2] or Schweizer [104, Thm. 16.14]. □

B.3 Measure theory

We only state the results we need for our work. For further reading on measure theory, we refer to the monographs of Diestel and Uhl [31] and Dinculeanu [32].

Proposition B.17 (Lebesgue theorem, [32, Thm II.§8.3.3], [11, Thm 6])**.** *Let X and E be Banach spaces with a bilinear mapping $X \times E \ni (x, e) \mapsto \langle x, e\rangle \in \mathbb{R}$ such that $|\langle x, e\rangle| \leq \|x\|_X \|e\|_E$ and let $\mu : \mathcal{B} \to E$ be a measure measure with finite variation. Let $(f^{(n)})_n$ be a sequence of μ-integrable X-valued functions on $[0, T]$ such that the $f^{(n)}$ converge to a function $f : [0, T] \to X$ μ-almost everywhere. Let $\|\mu\|$ denote the variation of μ. If a positive $\|\mu\|$-integrable function g exists such that $\|f^{(n)}(t)\|_X \leq g(t)$ for μ-almost all $t \in [0, T]$ and each $n \in \mathbb{N}$, then f is μ-integrable and $f^{(n)} \to f$, in particular*

$$\int_0^T f \,\mathrm{d}\mu = \lim_{n\to\infty} \int_0^T f^{(n)} \,\mathrm{d}\mu.$$

■

We found it non-trivial to extract Proposition B.17 from [32]. Therefore, we make the following clarifying remark.

Remark B.18. *Some of the prerequisites of* [32, Thm II.§8.3.3] *are stated on p. 119 at the beginning of Chap. II.§8. $\|\mu\|$ is a (inner and outer) regular Borel measure. Consequently, continuous functions like $t \mapsto \|f^{(n)}(t)\|_X$ are integrable w.r.t. the variation $\|\mu\|$, see* [32, Prop. III.§15.5.21 (p. 300)]*, which implies that a continuous function $f : [0, T] \to X$ is an element of the space $L^p((0, T), X, \mu)$ for the Borel measure μ. The oldest statement and proof of the result the author could find in the required generality is in Bartle's article* [11, Sect. 3, Thms 6 & 7]*.*

Proposition B.19 (Riesz–Markov–Kakutani theorem, [32, Chap. III.19, Cor. 2], [34])**.** *Let X be a Banach space. Then, there exists an isomorphism between the continuous linear functionals $\psi \in C([0, T], X)^*$ and regular Borel measures $\mu : \mathcal{B} \to X^*$ with finite variation defined by*

$$\psi(f) = \int_0^T f \,\mathrm{d}\mu$$

for $f \in C([0, T], X)$. ■

Definition B.20 (Absolutely continuous functions)**.** *Let X be a Banach space and I a compact interval. We call a function f with domain I and codomain X* ***absolutely continuous*** *if for each $\varepsilon > 0$, there exists $\delta > 0$ such that for all sequences $(a_n, b_n)_{n\in\mathbb{N}} \subset 2^I$ of pairwise disjoint sub-intervals with*

$$\sum_{n\in\mathbb{N}} (b_n - a_n) < \delta$$

we also have

$$\sum_{n\in\mathbb{N}} \|f(b_n) - f(a_n)\|_X < \varepsilon.$$

This definition of absolute continuity immediately implies uniform continuity.

Corollary B.21. *Let f be an absolutely continuous function. Then, it is uniformly continuous.* ■

From the theory of the Lebesgue integral, we have the following result.

Proposition B.22 ([88, Thm 2.5],[111, Prop. 1.12])**.** *Let I be a compact interval and $f \in L^1(I, X)$. Then, the mapping $t \mapsto \int_0^t f$ is absolutely continuous.* ■

The converse also holds for $X = \mathbb{R}$ by virtue of of the Radon-Nikodym theorem. However, this does not generalize to every Banach space. More specifically, requiring it to hold on a Banach space is equivalent to the Radon-Nikodym theorem holding on this Banach space, or, in other words, this Banach space having the Radon-Nikodym Property (RNP). We state some of Diestel's and Uhl's summary of equivalent formulations of the RNP [31].

Theorem B.23 ([31, Chap. VII.6])**.** *Let X be a Banach space. Then, the following are equivalent:*

1. *Consider a finite measure space $(\Omega, \mathcal{A}, \mu)$. For each μ-continuous vector measure $\lambda : \mathcal{A} \to X$ of bounded variation, there exists $f \in L^1(\Omega, X, \mu)$ such that*

   $$\lambda(A) = \int_A f \,\mathrm{d}\mu$$

   for all $A \in \mathcal{A}$. (Note: This is the statement of the Radon-Nikodym theorem.)

2. *The first property holds for every closed linear subspace of X.*

3. *All Bounded Variation (BV)-functions $f : I \to X$ are a.e. differentiable.*

4. *Every absolutely continuous function $f : I \to X$ is a.e. differentiable and we have*

   $$f(b) - f(a) = \int_a^b f'(x) \,\mathrm{d}x$$

   for all $a, b \in I$.

■

Definition B.24 (RNP)**.** *Let X be a Banach space. X is said to **have the RNP** if one of the equivalent statements in Theorem B.23 holds.*

Theorem B.25 (Rademacher's theorem [74, 92])**.** *Let X be a separable Banach space, Y be a Banach space that has the RNP. Consider a Lipschitz continuous mapping $f : G \to Y$ with $G \subset X$. Then, f is Frechet-differentiable almost everywhere.* ■

Remark B.26. *The statement of Theorem B.25 in the case of $X = \mathbb{R}$ is an equivalent characterization of Y having the RNP.*

Theorem B.27 ([31, Thm IV.1.1]). *Let (Ω, Σ, μ) be a finite measure space. Let $p \in [1, \infty)$ and $q \in (1, \infty]$ solve $\frac{1}{p} + \frac{1}{q} = 1$. Let X be a Banach space. Then, the following are equivalent:*

1. *The topological dual X^* has the RNP.*
2. *The identification $L^p(\Omega, X)^* = L^q(\Omega, X^*)$ holds true.*

■

Banach spaces that allow this identification are called Asplund spaces, which we state in the following definition

Definition B.28 (Asplund space). *Let X be a Banach space. Then, X is called an **Asplund space** if its topological dual X^* has the RNP.*

B.4 Properties of integrable functions

First, we assert that closing the class of smooth functions in the L^1-norm gives L^1.

Proposition B.29 ([80, Prop. B1]). *Let X be a Banach space. Then,*

$$\overline{C^\infty([0,T],X)}^{\|\cdot\|_{L^1((0,T),X)}} = L^1((0,T),X).$$

■

Next, we assert that the convolution of an L^1-function with a strongly continuous semigroup gives an L^1-function.

Proposition B.30 ([80, Prop. B2], [6, Prop. 1.3.4]). *Let X be a Banach space, $(S(t))_{t\geq 0}$ be a strongly continuous semigroup on X and $f \in L^1((0,T),X)$. Furthermore, let $0 < t \leq T$. Then, the mapping*

$$[0,t] \ni s \mapsto S(t-s)f(s) \in X$$

is measurable and integrable in the sense of Bochner, i.e. in $L^1((0,t),X)$. ■

Theorem B.31 (Grönwall's inequality [46] in integral form). *Let $a \in L^1((0,T),\mathbb{R})$, $b \in L^\infty((0,T),\mathbb{R})$ and $b(t) \geq 0$ for a.e. $t \in [0,T]$. If $x \in L^1((0,T),\mathbb{R})$ satisfies*

$$x(t) \leq a(t) + \int_0^t b(s)x(s)\,\mathrm{d}s$$

for a.a. $t \in [0,T]$. Then,

1. $x(t) \leq a(t) + \int_0^t a(s)b(s)\exp\left(\int_s^t b(\tau)\,\mathrm{d}\tau\right)\mathrm{d}s$,
2. *if* $a(s) \leq a(t)$ *for a.a.* $0 \leq s \leq t \leq T$, *we obtain*

$$x(t) \leq a(t)\exp\left(\int_0^t b(s)\,\mathrm{d}s\right)$$

for a.a. $t \in [0,T]$. *If* x *and* a *are continuous and the prerequisites hold everywhere on* $[0,T]$, *the claims hold everywhere on* $[0,T]$.

Proof. We convince ourselves that the mapping $[0,T] \ni t \mapsto \exp\left(-\int_0^t b(s)\,\mathrm{d}s\right) \in \mathbb{R}^+$ is in $C([0,T],\mathbb{R})$. Now, the reasoning in Teschl's textbook [116, Lem. 2.7] can be applied with all terms being well-defined when restricting to almost all $t \in [0,T]$ to obtain the first claim. For the second claim, the Fundamental Theorem of Calculus (FTC) gives

$$\int_0^t b(s)\exp\left(\int_s^t b(\tau)\,\mathrm{d}\tau\right)\mathrm{d}s = \left[-\exp\left(\int_s^t b(\tau)\,\mathrm{d}\tau\right)\right]_{s=0}^{s=t}$$

for all $t \in [0,T]$ and the claim follows with the a.e. monotonicity of a. □

B.5 Miscellaneous

Definition B.32 (Completely continuous mapping)**.** *Let* X *and* Y *be Banach spaces. A continuous mapping* $A : X \to Y$ *is called* ***completely continuous*** *if for all* $(x_n)_n \subset X$ *with* $x_n \rightharpoonup x \in X$ *in* $\sigma(X, X^*)$, *we have* $Ax_n \to Ax \in Y$ *in* $(Y, \|\cdot\|_Y)$.

Theorem B.33 (Carathéodory's theorem, see [10, 18])**.** *Let* $S \subset \mathbb{R}^d$, $x \in \operatorname{conv} S$. *Then, there exists* $\alpha_1,\ldots,\alpha_L \in \mathbb{R}$, $x_1,\ldots,x_L \in S$ *for some* $L < \infty$ *such that*

$$x = \sum_{i=1}^{L} \alpha_i x_i.$$

Definition B.34 (k-extension property, see [94, Def. 7.11])**.** *Let* $\Omega \subset \mathbb{R}^n$ *be a bounded domain. Let* $k \in \mathbb{N}$. *Then,* Ω *has the* k-***extension property*** *if there exists* $E \in \mathcal{L}(H^k(\Omega), H^k(\mathbb{R}^n))$ *such that*

$$Eu|_\Omega = u \text{ for all } u \in H^k(\Omega).$$

B.6 Proof of Theorem 3.12

This section shows how the existence and uniqueness for the solution of (3.4) follow from Assumptions 3.10 and 3.11. We pursue the arguments in [93], which make heavily use of a comparison principle. We will summarize and cut some parts and extend others to improve the accessibility of the argument. We note that the arguments and summaries in [33, 85] supported the understanding of the arguments in [93].

B.6.1 Preparations

We begin with a series of assumptions, definitions and statements for the operators and linear IVPs, which we just state and for the proof of which we refer to [93] and the references therein. When we get closer to the treatment of the nonlinear term, we provide summaries and details of the proof in [93]. To simplify the comparison to the corresponding statements in [93], we have borrowed its notation for the constants that occur in the norm estimates in this subsection. We note that we use the notational abbreviation $\mathcal{V} = H_0^1(\Omega)$ and $\mathcal{H} = L^2(\Omega)$ from Section 3.4. However, we hope that the notation helps to observe that large parts of the proof do not rely on these specific Hilbert space choices, but hold in much more generality.

Assumption B.35 (Elliptic operator)**.** *Let the operator A be defined as in Assumption 3.10. We assume further that A and a constant k_1 satisfy*

$$\sum_{i=1}^{d}\sum_{j=1}^{d}\int_\Omega a_{ij}(x)D_j y(x)D_i y(x)\,\mathrm{d}x + k_1\int_\Omega y^2(x)\,\mathrm{d}x \geq \frac{m_0}{2}\|y\|_{\mathcal{V}}^2 \tag{B.1}$$

for all $y \in \mathcal{V}$.

Definition B.36 ([93, Sect. 3.1])**.** *The operator $(\tilde{A}, D(\tilde{A}))$ is defined by*

$$\tilde{A}y \coloneqq Ay + k_1 y \quad \textit{for} \quad y \in D(\tilde{A}) \coloneqq \left\{y \in C^2(\bar{\Omega}) : y|_{\partial\Omega} = 0\right\}.$$

The operator A_ℓ is defined as the closure of $\tilde{A}$ in $L^\ell(\Omega)$.

Proposition B.37 ([94, Thm 11.3])**.** *Let an operator A and a constant k_1 satisfy* (B.1)*. Then, $\tilde{A}$ is of maximal $L^2((0,T),\mathcal{V}^*)$-regularity.* ■

Proposition B.38 ([93, Sect. 3.1])**.** *Let an operator A and a constant k_1 satisfy* (B.1)*. Then,*

1. *$(-A_\ell)$ generates a strongly continuous semigroup $(S_\ell(t))_{t\geq 0}$ which is analytic for $1 \leq \ell < \infty$ with domain $D(A_\ell) = \left\{y \in W^{2,\ell}(\Omega) : y|_{\partial\Omega} = 0\right\}$.*
2. *$0 \in \rho(-A_\ell)$*
3. *$A_\ell^\gamma \coloneqq (A_\ell^{-\gamma})^{-1}$*

■

For the semigroups $(S_\ell(t))_{t\geq 0}$, the following exponential estimates hold true.

Lemma B.39 ([93, Lem. 3.1])**.** *For all $1 \leq \ell \leq \lambda \leq \infty$ with $\ell < \infty$, there exists $C_2 > 0$ such that*

$$\|S_\ell(t)\varphi\|_{L^\lambda(\Omega)} \leq C_2 t^{-\frac{d}{2}\left(\frac{1}{\ell}-\frac{1}{\lambda}\right)}\|\varphi\|_{L^\ell(\Omega)}$$

for all $\varphi \in L^\ell(\Omega)$ *and* $t > 0$. *For all* $1 \le \ell \le \lambda \le \infty$ *with* $\ell < \infty$ *and all* $\alpha > 0$ *there exists* $C_3 > 0$ *such that*

$$\|A_\ell^\alpha S_\ell(t)\varphi\|_{L^\lambda(\Omega)} \le C_3 t^{-\frac{d}{2}(\frac{1}{\ell}-\frac{1}{\lambda})-\alpha}\|\varphi\|_{L^\ell(\Omega)}.$$

■

The following statements provide existence and uniqueness for solutions of linear IVPs with increasing complexity in the linear summand that is added on the left hand side.

Proposition B.40 ([93, Rem. 3.2]). *Let Assumption B.35 hold. Let* $y_0 \in L^\infty(\Omega)$ *and let* $f \in L^\mu((0,T), L^m(\Omega))$. *Then, there exists a unique weak solution* $y \in C([0,T],\mathcal{H}) \cap L^2((0,T),\mathcal{V})$ *of the IVP*

$$\partial_t y + Ay + k_1 y = f \text{ in } \Omega_T, \quad y|_{\partial\Omega} = 0, \quad y(0) = y_0 \tag{B.2}$$

■

Proposition B.41 ([93, Prop. 3.1]). *Let* $f \in L^\mu((0,T), L^m(\Omega))$, $y_0 \in L^\infty(\Omega)$. *Let* μ' *denote the Hölder conjugate of* μ. *Let the constants* μ *and* m *satisfy*

$$\mu > 1, \quad m > 1, \quad \frac{m}{\mu'} > \frac{d}{2}.$$

Let Assumption B.35 hold. Then, the unique weak solution of (B.2) *satisfies* $y \in L^\infty(\Omega_T) \cap C(\bar{\Omega}_{\varepsilon,T})$ *for every* $\varepsilon > 0$ *and the estimate*

$$\|y\|_{L^\infty(\Omega_T)} \le C_4(\|f\|_{L^\mu((0,T),L^m(\Omega))} + \|y_0\|_{L^\infty(\Omega)})$$

for some $C_4 > 0$. *In case* $\mu = m$, *we require* $\mu = m > \frac{d}{2} + 1$. *Moreover,* $y \in \mathcal{W}$. ■

Lemma B.42 ([93, Prop. 3.1]). *Let* $q > \frac{d}{2} + 1$. *Let* a, $f \in L^q(\Omega_T)$, $y_0 \in \mathcal{H}$. *Let* y *be the weak solution of*

$$\partial_t y + Ay + ay = f \text{ in } \Omega_T, \quad y|_{\partial\Omega} = 0, \quad y(0) = y_0. \tag{B.3}$$

in $C([0,T],\mathcal{H}) \cap L^2((0,T),\mathcal{V})$. *Then,* $[y]^+$ *satisfies*

$$\begin{aligned}\frac{1}{2}\int_\Omega ([y]^+(T)^2 + [y]^+(0)^2) + \sum_{i,j}\int_{\Omega_T} a_{ij} D_j [y]^+ D_i [y]^+ + \int_\Omega a([y]^+)^2 \\ = \int_\Omega y_0 [y]^+(0) + \int_{\Omega_T} f [y]^+.\end{aligned}$$

■

Definition B.43. *For* $a \in L^1(\Omega_T)$, *we say that* $((a_{ij})_{ij}, a)$ *satisfies the ellipticity condition* (Em_0) *if for a.e.*

$$\int_\Omega \sum_{i,j} a_{ij}(x) D_j\varphi(x) D_i\varphi(x)\,\mathrm{d}x + \int_\Omega a(t,x)\varphi(x)^2\,\mathrm{d}x \ge \frac{m_0}{2}\|\varphi\|_{\mathcal{V}}^2 \tag{Em_0}$$

for a.a. $t \in [0,T]$ *and for all* $\varphi \in \mathcal{V}$.

Proposition B.44 (Comparison principle, [93, Prop. 3.2]). *Let $q > \frac{d}{2}+1$, a, $f \in L^q(\Omega_T)$, $y_0 \in \mathcal{H}$, $a(t,x) \geq C_0$ in Ω_T and $y \in C([0,T],\mathcal{H}) \cap L^2((0,T),\mathcal{V})$ be the weak solution of* (B.3). *Let $f \leq 0$, $y_0 \leq 0$. Then, $y \leq 0$ in Ω_T and $y(T,\cdot) \leq 0$ in Ω.*

Proof. We summarize the proof from [93, Prop. 3.2].

Case $((a_{ij})_{ij}, a)$ satisfies (Em_0)**:** From $y_0 \leq 0$ and Lemma B.39, we obtain

$$0 \leq \frac{1}{2}\int_\Omega ([y]^+(T)^2 + [y]^+(0)^2) + \sum_{i,j}\int_{\Omega_T} a_{ij} D_j [y]^+ D_i [y]^+ + \int_{\Omega_T} a([y]^+)^2 \leq \int_{\Omega_T} f [y]^+ .$$

Using (Em_0) and $f \leq 0$, we obtain

$$0 \leq \frac{1}{2}\int_\Omega ([y]^+(T)^2 + [y]^+(0)^2) + \frac{m_0}{2}\| [y]^+ \|_{L^2((0,T),\mathcal{V})} \leq 0,$$

i.e. $[y]^+ \leq 0$ and $[y]^+(T) \leq 0$ and consequently, $y \leq 0$ and $y(T) \leq 0$.

Case $((a_{ij})_{ij}, a)$ need not satisfy (Em_0)**:** For $\theta \in \mathbb{R}$, we define $w(t,x) := \exp(-\theta t) y(t,x)$. Then, w is the weak solution of

$$\partial_t w + Aw + (a+\theta)w = f\exp(-\theta t) \text{ in } \Omega_T, \quad w|_{\partial\Omega} = 0, \quad w(0) = y_0.$$

Setting $\tilde{C} := \min\{0, C_0\}$ and let $\psi \in \mathcal{V}$ we obtain

$$\sum_{i,j}\int_\Omega a_{ij} D_j\psi D_i\psi + \int_\Omega (a(t)+\theta)\psi^2 \geq m_0\|D\psi\|^2_{\mathcal{H}} + (\tilde{C}+\theta)\|\psi\|^2_{\mathcal{H}}.$$

If $\tilde{C} = 0$, we set $\theta := m_0$. If $\tilde{C} < 0$, we set $\theta := \frac{m_0}{2} - \tilde{C}$. Both cases yield $\theta > 0$ and $((a_{ij})_{ij}, a+\theta)$ satisfying (Em_0). Now, we can apply the first case to w and obtain $w \leq 0$ and $w(T,\cdot) \leq 0$. The claim follows from $y(t,x) = \exp(t\theta)w(t,x)$. □

Proposition B.45 ([93, Prop. 3.3]). *Let $q > \frac{d}{2}+1$. Let $a \in L^q(\Omega_T)$ satisfy $a(t,x) \geq C_0$ in Ω_T. Then, there exists a constant $C_5 > 0$, independent of a but not of C_0, such that for every $f \in L^q(\Omega_T)$ and every $y_0 \in L^\infty(\Omega)$, the weak solution $y \in C([0,T],\mathcal{H}) \cap L^2((0,T),\mathcal{V})$ of* (B.3) *is an element of $L^\infty(\Omega_T) \cap C(\bar{\Omega}_{\varepsilon,T})$ for every $0 < \varepsilon < T$. In particular, the estimate*

$$\|y\|_{L^\infty(\Omega_T)} \leq C_5(\|f\|_{L^q(\Omega_T)} + \|y_0\|_{L^\infty(\Omega)})$$

holds and $y \in \mathcal{W}$.

Proof. We follow the proof of [93, Prop. 3.3] and consider the solutions y_1 (y_2) of (B.3) for $[f]^+$ and $[y_0]^+$ ($[f]^-$ and $[y_0]^-$). Proposition B.44 implies $y_1 \geq 0$, $y_2 \geq 0$ in Ω_T and $y := y_1 - y_2$ solves (B.3) for f and y_0. It remains to prove the estimate.

With the argument from the proof of Proposition B.44, there exists $\theta > 0$ such that $((a_{ij})_{ij}, C_0 + \theta)$ satisfies (Em_0) and the choice $k_1 = C_0 + \theta$ satisfies (B.1), i.e. Assumption B.35. Thus, we can consider w being the weak solution of the IVP

$$\partial_t w + Aw + (C_0 + \theta)w = [f]^+ \exp(-\theta t) \text{ in } \Omega_T, \quad w|_{\partial\Omega} = 0, \quad w(0) = [y_0]^+ .$$

Proposition B.44 implies $w \geq 0$ in Ω_T. Proposition B.41 and (Em_0) imply $w \in L^\infty(\Omega_T)$ and the estimate

$$\|w\|_{L^\infty(\Omega_T)} \leq K(\| [f]^+ \|_{L^q(\Omega_T)} + \| [y_0]^+ \|_{L^\infty(\Omega)}).$$

Setting $z := \exp(-\theta t)y_1 - w$ implies that z solves

$$\partial_t z + Az + (a + \theta)z = (C_0 - a)w \text{ in } \Omega_T, \quad z|_{\partial\Omega} = 0, \quad z(0) = 0.$$

We combine Proposition B.44 with the insight $(C_0 - a)w \leq 0$ and deduce $z \leq 0$ a.e. in Ω_T. Consequently, we estimate

$$0 \leq y_1 \leq \exp(\theta t)w \text{ a.e. in } \Omega_T$$

and

$$\|y_1\|_{L^\infty(\Omega_T)} \leq K(\| [f]^+ \|_{L^q(\Omega_T)} + \| [y_0]^+ \|_{L^\infty(\Omega)}).$$

Proposition B.41 implies $y_1 \in C(\bar{\Omega}_{\varepsilon,T})$ for every $\varepsilon > 0$. The estimate for y_2 is obtained in the same way and the claimed estimate follows from the triangle inequality. $y \in \mathcal{W}$ follows from the prerequisites, e.g. with [94, Thm 11.3]. □

Proposition B.46 ([93, Prop. 3.4])**.** *Let $q > \frac{d}{2} + 1$. Let $a \in L^q(\Omega_T)$ satisfy $a(t,x) \geq C_0$ in Ω_T. Let $0 < \varepsilon < T$. Then, there exists $C_6(\varepsilon) > 0$, independent of a but not of C_0, such that for every $f \in L^q(\Omega_T)$ and for every $y_0 \in L^\infty(\Omega)$, the weak solution $y \in \mathcal{W}$ of* (B.3) *is in $C(\bar{\Omega}_{\varepsilon,T})$ and satisfies*

$$\|y\|_{C(\bar{\Omega}_{\varepsilon,T})} \leq C_6(\varepsilon)(\|f\|_{L^q(\Omega_T)} + \|y_0\|_{\mathcal{H}}).$$

Proof. We follow the proof of [93, Prop. 3.4]. Having the proof of Proposition B.45 at hand, it remains to show that the weak solution w of the equation

$$\partial_t w + Aw + aw = [f]^{\mathfrak{s}} \text{ in } \Omega_T, \quad w|_{\partial\Omega} = 0, \quad w(0) = [y_0]^{\mathfrak{s}}$$

satisfies $\|w(t)\|_{L^\infty(\Omega)} \leq K(\varepsilon)(\|f\|_{L^q(\Omega_T)} + \|y_0\|_{\mathcal{H}})$ for some constant $K(\varepsilon) > 0$ for every $t \in [\varepsilon, T]$ and for $\mathfrak{s} \in \{+, -\}$. We consider the simpler IVPs

$$\partial_t \bar{w} + A\bar{w} + (C_0 + \theta)\bar{w} = [f]^{\mathfrak{s}} \text{ in } \Omega_T, \quad w|_{\partial\Omega} = 0, \quad w(0) = [y_0]^{\mathfrak{s}}$$

for $\mathfrak{s} \in \{+, -\}$ with $\theta > 0$ chosen to assert that $((a_{ij})_{ij}, C_0)$ satisfies (Em_0) and denote the solutions by $\bar{w}_1$ and $\bar{w}_2$. Lemma B.42 with the choices $\ell = 2$ and $\lambda = \infty$ and the variation of constants formula give

$$\|\bar{w}_i(t)\|_{L^\infty(\Omega)} \leq K\varepsilon^{-\frac{d}{4}}\|y_0\|_{\mathcal{H}} + K\| [f]^{\mathfrak{s}} \|_{L^q(\Omega_T)}$$

for some $K > 0$ for every $\varepsilon \leq t \leq T$ for $(i, \mathfrak{s}) \in \{(1,+),(2,-)\}$. Furthermore, Proposition B.44 implies $\bar{w}_1 \geq 0$ and $\bar{w}_2 \geq 0$. Let y_1, y_2 be as in the proof of Proposition B.45 as well as $\theta_i > 0$. With the settings $z_i \coloneqq \exp(-\theta_i t) y_i - \bar{w}_i$ for $i \in \{1,2\}$, we obtain that the z_i solve

$$\partial_t z_i + A z_i + (a + \theta_i) z_i = (C_0 - a)\bar{w}_i \quad , z|_{\partial\Omega} = 0, \quad z_i(0) = 0$$

with $z_i \leq 0$ and deduce $0 \leq y_i \leq \exp(\theta_i t)\bar{w}_i$ for $i \in \{1,2\}$ analogously to the proof of Proposition B.45. Combining the derived estimates yields the claim. □

As in [93, Appendix A], we give an assumption of a reduced setting for which existence, uniqueness and estimates are easier to establish.

Assumption B.47 ([93, Ass. (A1')]). *Let $q > \frac{d}{2} + 1$. Let $f^* : \Omega_T \times \mathbb{R} \to \mathbb{R}$ be given. For every $y \in \mathbb{R}$, let the function $f^*(\cdot, y) : \Omega_T \to \mathbb{R}$ be measurable. For a.a. $(t,x) \in \Omega_T$, let $f^*((t,x),\cdot) \in C^1(\mathbb{R})$ and let the estimates*

$$\begin{aligned} |f^*((t,x),0)| &\leq M_1^*((t,x)) \\ C_0 \leq \partial_y f^*((t,x),y) &\leq M_1^*((t,x)) \end{aligned}$$

hold with a function $M_1^* \in L^{2q}(\Omega_T)$ and a constant $C_0 \in \mathbb{R}$.

The existence and uniqueness claim under Assumption B.47 is stated below. Again, we follow the ideas from [93], but give more details to improve the accessibility of the argument in [93, Lem. A.1].

Lemma B.48 ([93, Lem. A.1]). *Let $q > \frac{d}{2} + 1$. Let Assumption B.47 hold. Let $y_0 \in L^\infty(\Omega)$. Then, the IVP*

$$\partial_t y + A y + f^*((t,x),y) = 0 \text{ in } \Omega_T, \quad y|_{\partial\Omega} = 0, \quad y(0) = y_0 \tag{B.4}$$

admits a unique weak solution $y \in \mathcal{W} \cap L^\infty(\Omega_T)$, which satisfies the estimate

$$\|y\|_{L^\infty(\Omega_T)} + \|y(T)\|_{L^\infty(\Omega)} + \|y\|_{L^2((0,T),\mathcal{V})} \leq C_{11}(\|f^*(\cdot,0)\|_{L^q(\Omega_T)} + \|y_0\|_{L^\infty(\Omega)})$$

for some constant $C_{11} > 0$.

Proof. We keep with the ideas from [93, Lem. A.1]. If $((a_{ij})_{ij}, C_0)$ does not satisfy (Em$_0$), we consider $w(t,x) \coloneqq \exp(-\theta t) y(t,x)$ with $\theta > 0$ such that $((a_{ij})_{ij}, C_0 + \theta)$ satisfies (Em$_0$) and note that y solves (B.4) weakly if and only if w solves

$$\partial_t w + A w + \tilde{f}^*((t,x),w) = 0 \text{ in } \Omega_T, \quad w(0) = y_0$$

weakly with $\tilde{f}^*((t,x),w) = \exp(-\theta t) f^*((t,x)\exp(\theta t)w) + \theta w$. Furthermore, the estimate $\partial_y \tilde{f}^*((t,x),w) \geq C_0 + \theta$ follows immediately, which yields that we can assume without loss of generality that $((a_{ij})_{ij}, C_0)$ satisfies (Em$_0$). As noted in [93], we note

that the existence of a weak solution $y \in \mathcal{W}$ follows along the lines of [75, Chap. 2, Thm 1.2]. Therein, the statement is proven for a class nonlinear monotone operators, which is matched by our setting, in which the linear operator A is perturbed by the superposition operator induced by f^*. As the derivative $\partial_y f^*$ is bounded from below, the perturbed operator can be designed to satisfy the required monotonicity. In particular, y solves the IVP

$$\partial_t w + Aw + a^* w = -f^*((t,x),0) \text{ in } \Omega_T, \quad w|_{\partial\Omega} = 0, \quad w(0) = y_0 \text{ in } \Omega_T \tag{B.5}$$

with $a^*((t,x)) = \int_0^1 \partial_y \tilde{f}^*((t,x),\theta y)\,\mathrm{d}\theta \geq C_0$ as well. Propositions B.45 and B.46 yield $y \in C(\bar{\Omega}_{\varepsilon,T}) \cap L^\infty(\Omega)$ and

$$\|y\|_{L^\infty(\Omega_T)} + \|y(T)\|_{L^\infty(\Omega)} \leq K(\|f^*(\cdot,0)\|_{L^q(\Omega_T)} + \|y_0\|_{L^\infty(\Omega)}).$$

The estimate in $L^2((0,T),\mathcal{V})$ can be derived as follows. We test (B.5) with the solution y and integrate over $[0,T]$ to get the identity

$$\int_0^T \langle \partial_t y, y\rangle_{\mathcal{V}^*,\mathcal{V}} + \int_0^T \sum_{ij} \langle a_{ij} D_i y, D_j y\rangle_{\mathcal{H}} + \int_0^T a^* \langle y, y\rangle_{\mathcal{H}} = \int_0^T \langle f^*(0), y\rangle_{\mathcal{V}^*,\mathcal{V}}\,.$$

We apply (Em_0) and the Cauchy–Schwarz inequality, which give

$$\int_0^T \langle \partial_t y, y\rangle_{\mathcal{V}^*,\mathcal{V}} + \frac{m_0}{2}\|y\|^2_{L^2((0,T),\mathcal{V})} \leq \int_0^T \|f^*(0)\|_{\mathcal{V}^*}\|y\|_{\mathcal{V}}.$$

Furthermore, integration by parts gives $\int_0^T \langle \partial_t y(t), y(t)\rangle_{\mathcal{V}^*,\mathcal{V}}\,\mathrm{d}t = \frac{1}{2}\,\|y(t)\|^2_{\mathcal{H}}|_{t=0}^{t=T}$. Plugging this into the estimate above yields

$$\frac{1}{2}\left(\|y(T)\|^2_{\mathcal{H}} - \|y_0\|^2_{\mathcal{H}}\right) + \frac{m_0}{2}\|y\|^2_{L^2((0,T),\mathcal{V})} \leq \int_0^T \|f^*((t,\cdot),0)\|_{\mathcal{V}^*}\|y(t)\|_{\mathcal{V}}\,\mathrm{d}t.$$

We use the inequality of arithmetic and geometric means, more precisely

$$\|f^*((t,\cdot),0)\|_{\mathcal{V}^*}\sqrt{\frac{2}{m_0}}\sqrt{\frac{m_0}{2}}\|y(t)\|_{\mathcal{V}} \leq \frac{1}{2}\left(\frac{2}{m_0}\|f^*((t,\cdot),0)\|^2_{\mathcal{V}^*} + \frac{m_0}{2}\|y(t)\|^2_{\mathcal{V}}\right),$$

to obtain

$$\frac{1}{2}\left(\|y(T)\|^2_{\mathcal{H}} - \|y_0\|^2_{\mathcal{H}}\right) + \frac{m_0}{4}\|y\|^2_{L^2((0,T),\mathcal{V})} \leq \frac{1}{m_0}\|f^*(0)\|^2_{L^2((0,T),\mathcal{V}^*)}\,\mathrm{d}t$$

and consequently,

$$\|y\|^2_{L^2((0,T),\mathcal{V})} \leq C\left(\|f^*(0)\|^2_{L^2((0,T),\mathcal{V}^*)} + \|y_0\|^2_{\mathcal{H}}\right)$$

for some $C > 0$, which yields the claim by using that the $\sqrt{\cdot}$-function is positive subadditive, the continuous embeddings $L^q(\Omega_T) \hookrightarrow L^2(\Omega_T) \cong L^2((0,T),\mathcal{H}) \hookrightarrow L^2((0,T),\mathcal{V}^*)$ hold and the embedding $L^\infty(\Omega) \hookrightarrow \mathcal{H}$. A short version of this argument can be found in several textbooks on the theory of parabolic equations, we refer to [26, Chap. XVIII, Sect. §3.3.2]. □

B.6.2 Main argument

After this preparatory work, we are able to prove Theorem 3.12. We note that this proof sets a different focus than the one in [93]. We refer to proofs of the preparatory statements and the article [93] for parts, which we find easily accessible and provide more details on the truncation and regularization technique that is used to reduce the problem to easier classes of problems.

Proof. The uniqueness follows from the comparison principle in Proposition B.44, see [93, Thm 3.1]. Applying the same reasoning as in the proof of Lemma B.48 allows us to assume that $((a_{ij})_{ij}, C_0)$ satisfies (Em_0) without loss of generality. As in the proof of Theorem 9.9, we commit a little abuse of notation and abbreviate $f(y,z) \coloneqq \sum_{i=1}^{M} \alpha_i f_i^a(y,v_i) + f^b(y,u)$ for the abbreviation $z = (u,\alpha) \in \mathbb{R} \times \mathbb{R}^M$. This abuse of notation is conducted safely as we never use the meaning given by (MIPEVO-TX) and Assumption 3.11 in this proof.

We continue as in [93, Part of the proof of Thm 3.1 in the Appendix]. Regarding the structure of the remaining parts of the proof, we begin by introducing a truncation and mollification of the function f and obtain regularity properties for them. Then, we consider the family of IVPs for the truncated and regularized nonlinear terms and employ the preparations to obtain existence of weak solutions and estimates on their norms. Then, we give reformulation of the arising IVPs with truncated and mollified right hand sides by virtue of a corresponding sequence of monotone operators. Next, we deduce that the sequence of mollified solutions for a fixed truncation has a weak limit. Then, it is shown that the weak limit point indeed solves the truncated state equation and in turn state equation (3.4) if the truncation value is chosen large enough. The argument closes with a reduction of the problem to the setting of Proposition B.46, which gives the desired norm estimates.

Truncation. Let $k \in \mathbb{N} \cup \{0\}$. Analogously to [93, p. 173], we define the linear truncation f_k of the function f by

$$f_k(y,z) \coloneqq \begin{cases} f(k,z) + \partial_y f(k,z)(y-k) & \text{if } y > k \\ f(y,z) & \text{if } |y| \leq k \\ f(-k,z) + \partial_y f(-k,z)(y+k) & \text{if } y < -k \end{cases}$$

for $y \in \mathbb{R}$ and $z \in \mathbb{R}^{1+M}$. We make the following observation with respect to the regularity of the truncations. Let $z \in \mathbb{R}^{1+M}$ be fixed. Then, Assumption 3.11 implies that the functions $f_k : \mathbb{R} \times \mathbb{R}^{1+M} \to \mathbb{R}$ are Lipschitz continuous in the first argument in the interval $[-k,k]$, i.e.

$$\max_{y \in [-k,k]} \partial_y f_k(y,z) \leq (M_1 + \max\{m_1, M_1\}\|z\|_1)\eta(|k|).$$

Furthermore, they are Lipschitz continuous in $[k,\infty)$ and $(-\infty,k]$ as well with Lipschitz constants $|\partial_y f(k,z)| \leq (M_1 + \max\{m_1,M_1\}\|z\|_1)\eta(|k|)$ and $|\partial_y f(-k,z)| \leq (M_1 +$

$\max\{m_1, M_1\}\|z\|_1)\eta(|k|)$ and thus, globally Lipschitz continuous. We consider the superposition operators

$$\begin{aligned} f_k(\cdot, z) : L^{q'}(\Omega_T) &\to L^q(\Omega_T) \hookrightarrow L^2(\Omega_T) \\ y &\mapsto f(y, z) \end{aligned}$$

with the Hölder conjugate q' of q, in particular $q' < 2$. With the argued Lipschitz continuity, these superposition operators satisfy the Carathéodory conditions and are continuous accordingly, see e.g. [94, Thm 10.58]. Moreover, the Lipschitz continuity transfers to the superposition operators if they considered as mappings between L^∞ spaces $f_k(\cdot, z) : L^\infty(\Omega_T) \to L^\infty(\Omega_T)$, which can be deduced from the chain of estimates

$$\begin{aligned} &\|f_k(y_1, z) - f_k(y_2, z)\|^2_{L^\infty(\Omega_T)} \\ &\quad = \int_{\Omega_T} (f_k(y_1(t,x), z(t,x)) - f_k(y_2(t,x), z(t,x)))^2 \,\mathrm{d}(t,x) \\ &\quad \underset{\text{Lip.}}{\leq} \int_{\Omega_T} ((M_1 + \max\{m_1, M_1\}\|z(t,x)\|_1)|y_1(t,x) - y_2(t,x)|)^2 \,\mathrm{d}(t,x) \\ &\quad \underset{\text{Hölder}}{\leq} (2M_1^2\lambda(\Omega_T) + 2\max\{m_1, M_1\}^2\|z\|^2_{L^2(\Omega_T, \mathbb{R}^{1+M})})\|y_1 - y_2\|^2_{L^\infty(\Omega_T)} \end{aligned}$$

for two vectors $y_1,\ y_2 \in L^\infty(\Omega_T)$. Here, $L^2(\Omega_T, \mathbb{R}^M)$ is considered as equipped with the sum-norm topology in the codomain. However, as norms are equivalent on $\mathbb{R}^{1+M}$, this is no restriction as we can exchange this norm by cost of another constant factor in the last estimate.

Mollification. Let $z \in \mathbb{R}^{1+M}$. Then, for $k \geq 1$, the mapping $f_k(\cdot, z) : \mathbb{R} \to \mathbb{R}$ is not necessarily in the class C^1 anymore. We circumvent this and regularize it. To this end, we introduce the sequence $(f_k^{(n)})_n$ with the setting

$$f_k^{(n)}((t,x), y) \coloneqq (\theta^{(n)} * f_k(y, z(\cdot)))(t,x),$$

where $(\theta^{(n)})_n$ denotes a family of positive mollifiers, see [20, Sect. 2.6], and $z \in L^q(\Omega_T, \mathbb{R}^{1+M})$ the joint control input, see [93, p. 173]. As the mollification for realization $y = 0$, $f_k^{(n)}(\cdot, 0)$ does not depend on k, we introduce the notation $f^{(n)}(\cdot, 0) \coloneqq f_k{}^{(n)}(\cdot, 0) = \theta^{(n)} * f(0, z)$ for all k. Again, the Carathéodory conditions are satisfied and the superposition operator $y \mapsto f_k^{(n)}(\cdot, y(\cdot))$ is in the class $C(L^2(\Omega_T), L^2(\Omega_T))$.

Solving the truncated and regularized IVP. As in [93, p. 173], we consider the IVP with truncated and regularized nonlinear term:

$$\partial_t y + Ay + f_k^{(n)}((t,x), y) = 0 \text{ in } \Omega_T, \quad y|_{\partial\Omega} = 0, \quad y(0) = y_0. \tag{B.6}$$

We notice that the truncated and regularized function $f_k^{(n)}$ satisfies Assumption B.47, see [93, p. 174]. We refer to [20, Thm 2.6-1] for information on the improved regularity

due to the mollification. Consequently, for every k and $n \in \mathbb{N}$, Lemma B.48 yields the existence of a unique weak solution $y_k^{(n)} \in \mathcal{W}$ and there exist constants $K > 0$ and $K_1 > 0$ such that we can estimate

$$\begin{aligned}
&\|y_k^{(n)}\|_{L^\infty(\Omega_T)} + \|y_k^{(n)}(T)\|_{L^\infty(\Omega)} + \|y_k^{(n)}\|_{L^2((0,T),\mathcal{V})} \\
&\quad \underset{\text{Lemma B.48}}{\leq} K(\|f^{(n)}(\cdot,0)\|_{L^q(\Omega_T)} + \|y_0\|_{L^\infty(\Omega)}) \\
&\quad \leq K(\|f^{(n)}(\cdot,0) - f(0,z)\|_{L^q(\Omega_T)} + \|f(0,z)\|_{L^q(\Omega_T)} + \|y_0\|_{L^\infty(\Omega)}) \\
&\quad \underset{\text{Young}}{\leq} K_1(1 + \|f(0,z)\|_{L^q(\Omega_T)} + \|y_0\|_{L^\infty(\Omega)}),
\end{aligned} \tag{B.7}$$

see [93, p. 174]. We note that the third inequality holds by virtue of Young's inequality for convolutions and the properties of the mollifier family.

Monotone operator reformulation of the truncated and regularized IVPs. Let k and $n \in \mathbb{N}$ be given. We define $\mathcal{A}_k^{(n)} : L^2((0,T),\mathcal{V}) \to L^2((0,T),\mathcal{V}^*)$ as

$$\begin{aligned}
&\left\langle \mathcal{A}_k^{(n)}(w), \varphi \right\rangle_{L^2((0,T),\mathcal{V}^*),L^2((0,T),\mathcal{V})} \\
&\qquad := \sum_{ij} \int_{\Omega_T} a_{ij} D_j w D_i \varphi + \int_{\Omega_T} (f_k^{(n)}(\cdot,w) - f^{(n)}(\cdot,0))\varphi
\end{aligned}$$

for w and $\varphi \in L^2((0,T),\mathcal{V})$. Then, the fact that $((a_{ij})_{ij}, C_0)$ satisfies (Em_0) implies that $\mathcal{A}_k^{(n)}$ is well-defined and monotone for all k and $n \in \mathbb{N}$. Indeed, we plug in the definitions into the second summand and obtain the identities

$$\begin{aligned}
&\int_{\Omega_T} (f_k^{(n)}(\cdot,w) - f^{(n)}(\cdot,0))\varphi \\
&\qquad = \int_{\Omega_T}\int_{\Omega_T} \theta^{(n)}(s-\sigma)(f_k(w(s),z(\sigma)) - f_k(0,z(\sigma)))\,\mathrm{d}\sigma\varphi(s)\,\mathrm{d}s \\
&\qquad \underset{\text{FTC}}{=} \int_{\Omega_T}\int_{\Omega_T} \theta^{(n)}(s-\sigma)\int_0^1 \partial_y f_k(\theta w(s), z(\sigma)) w(s)\,\mathrm{d}\theta\,\mathrm{d}\sigma\,\varphi(s)\,\mathrm{d}s \\
&\qquad = \int_{\Omega_T}\int_{\Omega_T} \theta^{(n)}(s-\sigma)\int_0^1 \partial_y f_k(\theta w(s), z(\sigma))\,\mathrm{d}\theta\,\mathrm{d}\sigma\, w(s)\varphi(s)\,\mathrm{d}s.
\end{aligned}$$

As the mollifier functions $\theta^{(n)}$ are positive, we have

$$C_0 \leq \int \theta^{(n)}(s-\sigma)\int_0^1 \partial_y f_k(\theta w(s), z(\sigma))\,\mathrm{d}\theta\,\mathrm{d}\sigma,$$

which establishes (Em_0) as well as the monotony. We can use the first identity to obtain

$$\int_{\Omega_T} (f_k^{(n)}(\cdot,w) - f^{(n)}(\cdot,0))\varphi \underset{\text{C.S.}}{\leq} \|(f_k^{(n)}(\cdot,w) - f^{(n)}(\cdot,0))\|_{L^2(\Omega_T)}\|\varphi\|_{L^2(\Omega_T)}.$$

Regarding the first factor, we see

$$\begin{aligned}\int_{\Omega_T} &\theta^{(n)}(s-\sigma)(f_k(w(s),z(\sigma))-f_k(0,z(\sigma)))\,\mathrm{d}\sigma\\ &\leq \int_{\Omega_T}\theta^{(n)}(s-\sigma)|f_k(w(s),z(\sigma))-f_k(0,z(\sigma))|\,\mathrm{d}\sigma\\ &\leq \int_{\Omega_T}\theta^{(n)}(s-\sigma)(M_1+m_1\|z(\sigma)\|)\,\mathrm{d}\sigma|w(s)|\end{aligned}$$

If we additionally assume $w\in L^\infty(\Omega_T)$, we obtain

$$\|f_k^{(n)}(\cdot,w)-f^{(n)}(\cdot,0)\|_{L^2(\Omega_T)}\leq K_2\|w\|_{L^\infty(\Omega_T)}\|\varphi\|_{L^2(\Omega_T)} \tag{B.8}$$

for some constant $K_2>0$. Using the fact that the ellipticity condition (Em_0) holds, we can rewrite (B.6) as

$$\partial_t y+\mathcal{A}_k^{(n)}(y)=-f^{(n)}(\cdot,0)\text{ in }\Omega_T,\quad y|_{\partial\Omega_T}=0,\quad y(0)=y_0$$

because $z\in L^q(\Omega_T)\hookrightarrow L^2(\Omega_T)$.

We note that in [93, p. 174], the operator $\mathcal{A}_k^{(n)}$ is defined slightly different and also claimed to be monotone. The definition in [93, p. 174] simplifies the reformulation as the right hand side is zero, which leads to a less bloated chain of arguments. However, we have not managed to verify the monotonicity for the definition in [93, p. 174] and have therefore changed the definition to one, for which we were able to do so.

Existence of weak limits. We deviate in the derivation of weak limits. Let $\langle\cdot,\cdot\rangle$ denote the mapping that puts $L^2((0,T),\mathcal{V})$ and $L^2((0,T),\mathcal{V}^*)$ in duality. Let $k\in\mathbb{N}$. We estimate

$$\begin{aligned}&\sup_{\|\varphi\|\leq 1}\langle\mathcal{A}_k^{(n)}(y_k^{(n)}),\varphi\rangle\\ &\quad\leq\sup_{\|\varphi\|\leq1}\sum_{ij}\langle a_{ij}D_jy_k^{(n)},D_i\varphi\rangle+\sup_{\|\varphi\|\leq1}\int_{\Omega_T}(f_k^{(n)}(\cdot,y_k^{(n)})-f^{(n)}(\cdot,0))\varphi\\ &\quad\underset{\text{C.S. \& (B.8)}}{\leq}K_3\left(\int_0^T(\sum_j\|D_jy_k^{(n)}\|_{\mathcal{H}})^2\right)^{\frac12}\|\varphi\|_{L^2((0,T),\mathcal{V}^*)}+K_2\|y_k^{(n)}\|_{L^\infty(\Omega_T)}\|\varphi\|_{L^2(\Omega_T)}\\ &\quad\leq K_4\|y_k^{(n)}\|_{L^2((0,T),\mathcal{V})}+K_5\|y_k^{(n)}\|_{L^\infty(\Omega_T)},\end{aligned}$$

with constants K_3, K_4, $K_5>0$. We deduce boundedness of the sequence $(\mathcal{A}_k^{(n)}(y_k^{(n)}))_n$ in $L^2((0,T),\mathcal{V}^*)$ because the sequence $(y_k^{(n)})_n$ is bounded in $L^2((0,T),\mathcal{V})$ and $L^\infty(\Omega_T)$ by virtue of (B.7). As $\partial_t y_k^{(n)}=-\mathcal{A}_k^{(n)}(y_k^{(n)})-f^{(n)}(0,z)$, this boundedness implies the boundedness of the sequence $(y_k^{(n)})_n$ in $\mathcal{W}$ and $L^\infty(\Omega_T)$. Furthermore, the sequence $(y_k^{(n)}(T))_n$ is bounded in $\mathcal{H}$ and $L^\infty(\Omega)$. We employ the Banach-Alaoglu theorem to

deduce the existence of weak(*) limits $\bar{y}_k \in \mathcal{W} \cap L^\infty(\Omega_T)$ and $\chi_k \in L^2((0,T),\mathcal{V}^*)$ and a subsequence indexed by n' that satisfy

$$\begin{aligned} y_k^{(n')} &\rightharpoonup \bar{y}_k && \text{in } \mathcal{W}, \\ y_k^{(n')} &\rightharpoonup^* \bar{y}_k && \text{in } L^\infty(\Omega_T), \\ \mathcal{A}_k^{(n')}(y_k^{(n')}) &\rightharpoonup \chi_k && \text{in } L^2((0,T),\mathcal{V}^*), \\ y_k^{(n')}(T) &\rightharpoonup \bar{y}_k(T) && \text{in } \mathcal{H}. \end{aligned}$$

As the $y_k^{(n')}$ solve the truncated and mollified IVPs (B.6), we have

$$\left\langle \partial_t y_k^{(n')} + \mathcal{A}_k^{(n')}(y_k^{(n')}) + f^{(n')}(\cdot,0), \varphi \right\rangle = 0 \tag{B.9}$$

for all $\varphi \in L^2((0,T),\mathcal{V})$. We can pass to the limit $n' \to \infty$, which yields

$$\partial_t \bar{y}_k + f(0,z) = -\chi_k. \tag{B.10}$$

The weak limits solve the state equation Let $\langle \cdot,\cdot \rangle$ denote the mapping that puts $L^2((0,T),\mathcal{V})$ and $L^2((0,T),\mathcal{V}^*)$ in duality and define $\mathcal{A}_k$ by

$$\langle \mathcal{A}_k(w), \varphi \rangle \coloneqq \sum_{ij} \int_{\Omega_T} a_{ij} D_j w D_i \varphi + \int_{\Omega_T} (f_k(\cdot,w) - f(0,z))\varphi$$

for $w \in L^2((0,T),\mathcal{V}) \cap L^\infty(\Omega_T)$ and $\varphi \in L^2((0,T),\mathcal{V})$. If we are able to show that the identity $\chi_k = \mathcal{A}_k(\bar{y}_k)$ holds, we know that $\bar{y}$ solves

$$\partial_t y + Ay + f_k(y,z) = 0 \text{ in } \Omega_T, \quad y|_{\partial\Omega} = 0, \quad y(0) = y_0, \tag{B.11}$$

see [93, p. 175]. Due to the approximation of the family of mollifiers in $L^q(\Omega_T)$, we have $f_k^{(n')}(\cdot,w) \to f_k(w,z)$ in $L^q(\Omega_T)$ for fixed w and k, and consequently, $f_k^{(n')}(\cdot,w) - f^{(n')}(\cdot,0) \to f_k(\cdot,w) - f(0,z)$. In turn, we deduce that $\mathcal{A}_k^{(n')}(w) \to \mathcal{A}_k(w)$ in $L^2((0,T),\mathcal{V}^*)$ for a fixed $w \in L^2((0,T),\mathcal{V}) \cap L^\infty(\Omega_T)$. The monotony of $\mathcal{A}_k^{(n')}$ gives the estimate

$$\langle \mathcal{A}_k^{(n')}(w) - \mathcal{A}_k^{(n')}(y_k^{(n)}), w - y_k^{(n')} \rangle \geq 0$$

for any $w \in L^2((0,T),\mathcal{V}) \cap L^\infty(\Omega_T)$. We deduce the following equivalences

$$\begin{aligned} & 0 \leq \langle \mathcal{A}_k^{(n')}(w) - \mathcal{A}_k^{(n')}(y_k^{(n')}), w - y_k^{(n')} \rangle \\ \Leftrightarrow\ & 0 \leq \langle \mathcal{A}_k^{(n')}(w), w - y_k^{(n')} \rangle - \langle \mathcal{A}_k^{(n')}(y_k^{(n')}), w \rangle - \langle \partial_t y_k^{(n')}, y_k^{(n')} \rangle - \langle f^{(n')}(\cdot,0), y_k^{(n)} \rangle \\ \Leftrightarrow\ & 0 \leq \langle \mathcal{A}_k^{(n')}(w), w - y_k^{(n')} \rangle - \langle \mathcal{A}_k^{(n')}(y_k^{(n')}), w \rangle - \frac{1}{2}\|y_k^{(n')}(T)\|_{\mathcal{H}}^2 + \frac{1}{2}\|y_0\|_{\mathcal{H}}^2 \\ & \quad - \langle f^{(n')}(\cdot,0), y_k^{(n')} \rangle \end{aligned}$$

from (B.9) and integration by parts. The above convergence properties, the fact that the duality pairing of a norm convergent and a weakly convergent sequence converges and the fact that the norm is a weakly lower semi-continuous function yield

$$0 \leq \langle \mathcal{A}_k(w), w - y_k \rangle - \langle \chi_k, w \rangle - \frac{1}{2}\|\bar{y}_k(T)\|_{\mathcal{H}}^2 + \frac{1}{2}\|y_0\|_{\mathcal{H}}^2 - \langle f(0,z), \bar{y}_k \rangle$$

when passing to the limit $n' \to \infty$. We insert a zero and get an equivalence to

$$0 \leq \langle \mathcal{A}_k(w) - \chi_k, w - \bar{y}_k \rangle - \langle \chi_k, \bar{y}_k \rangle - \frac{1}{2}\|\bar{y}_k(T)\|_{\mathcal{H}}^2 + \frac{1}{2}\|y_0\|_{\mathcal{H}}^2 - \langle f(0,z), \bar{y}_k \rangle.$$

Plugging in the identity (B.10) into the previous estimate gives

$$\begin{aligned} 0 &\leq \langle \mathcal{A}_k(w) - \chi_k, w - \bar{y}_k \rangle + \langle \partial_t \bar{y}_k, \bar{y}_k \rangle - \frac{1}{2}\|\bar{y}_k(T)\|_{\mathcal{H}}^2 + \frac{1}{2}\|y_0\|_{\mathcal{H}}^2 \\ &\leq \langle \mathcal{A}_k(w) - \chi_k, w - \bar{y}_k \rangle. \end{aligned}$$

As in [93, p. 175 f.] we note that the estimate holds for the particular choice $w = \bar{y}_k - \lambda\psi$ with $\lambda > 0$ and $\psi \in L^2((0,T),\mathcal{V}) \cap L^\infty(\Omega_T)$ as w is arbitrary, which yields

$$\langle \mathcal{A}_k(\bar{y}_k - \lambda\psi) - \chi_k, \psi \rangle \leq 0.$$

The Lipschitz continuity of $f_k(\cdot, z) : L^\infty(\Omega_T) \to L^\infty(\Omega_T)$ implies that we can pass to the limit $\lambda \to 0$ in the equation above and obtain

$$\langle \mathcal{A}_k(\bar{y}_k) - \chi_k, \psi \rangle \leq 0$$

for every $\psi \in L^2((0,T),\mathcal{V}) \cap L^\infty(\Omega_T)$, which gives $\chi_k = \mathcal{A}_k(\bar{y}_k)$ in $L^2((0,T),\mathcal{V}^*)$, see [93, p. 176]. Consequently, the vector $\bar{y}_k$ solves the truncated IVP (B.11) and satisfies the estimate (B.7). We choose $k_0 \geq K_1(1 + \|y_0\|_{L^\infty(\Omega_T)} + \|f(0,z)\|_{L^q(\Omega_T)})$ and obtain the identity $f_{k_0}(\bar{y}_{k_0}, z) = f(\bar{y}_{k_0}, z)$ and the fact $\bar{y}_{k_0}$ solves (3.4), which is what we set out for, see [93, p. 176].

Norm estimates We finish with a bootstrapping argument: if y solves the IVP (3.4), it also solves

$$\partial_t w + Aw + aw = -f(0,z) \text{ in } \Omega_T, \quad w|_{\partial\Omega} = 0, \quad w(0) = y_0$$

if we define $a(t,x) \coloneqq \int_0^1 \partial_y f(\theta y, z)\,\mathrm{d}\theta \geq C_0$. Here, the boundedness from below stems from Assumption 3.11. Furthermore, we remember that $-f(0,z) \in L^q(\Omega_T)$ and can apply Proposition B.46 to establish the desired norm estimates, see [93, p. 176]. □

Bibliography

[1] W. Achtziger and C. Kanzow, *Mathematical programs with vanishing constraints: optimality conditions and constraint qualifications*, Mathematical Programming **114** (2008), no. 1, 69–99. doi:10.1007/s10107-006-0083-3.

[2] R. A. Adams, *Sobolev spaces*, Academic Press, New York, 1975.

[3] M. S. Alnæs, J. Blechta, J. Hake, A. Johansson, B. Kehlet, A. Logg, C. Richardson, J. Ring, M. E. Rognes, and G. N. Wells, *The FEniCS Project Version 1.5*, Archive of Numerical Software **3** (2015), no. 100. doi:10.11588/ans.2015.100.20553.

[4] H. Antil and J. Pfefferer, *A short Matlab implementation of fractional Poisson equation with nonzero boundary conditions*, Technical Report, 2017.

[5] T. Apel and G. Lube, *Anisotropic mesh refinement in stabilized Galerkin methods*, Numerische Mathematik **74** (1996), no. 3, 261–282. doi:10.1007/s002110050216.

[6] W. Arendt, C. J. K. Batty, M. Hieber, and F. Neubrander, *Vector-valued Laplace transforms and Cauchy problems*, Vol. 96, Springer Science & Business Media, 2011. doi:10.1007/978-3-0348-0087-7.

[7] Z. Artstein, *Yet another proof of the Lyapunov convexity theorem*, Proceedings of the American Mathematical Society (1990), 89–91. doi:10.2307/2047697.

[8] J.-P. Aubin and A. Cellina, *Differential inclusions: set-valued maps and viability theory* (1984). doi:10.1007/978-3-642-69512-4.

[9] F. Bachmann, D. Beermann, J. Lu, and S. Volkwein, *Pod-based mixed-integer optimal control of the heat equation*, Journal of Scientific Computing (2019), 1–28. doi:10.1007/s10915-019-00924-3.

[10] I. Bárány, *A generalization of Carathéodory's theorem*, Discrete Mathematics **40** (1982), no. 2-3, 141–152. doi:10.1016/0012-365X(82)90115-7.

[11] R. Bartle, *A general bilinear vector integral*, Studia Mathematica **15** (1955), no. 3, 337–352. doi:10.4064/sm-15-3-337-352.

[12] M. P. Bendsoe and O. Sigmund, *Topology optimization – theory, methods and applications*, Springer, 2003. doi:10.1007/978-3-662-05086-6.

[13] H. G. Bock, C. Kirches, A. Meyer, and A. Potschka, *Numerical solution of optimal control problems with explicit and implicit switches*, Optimization Methods and Software (2018), 1–25. doi:10.1080/10556788.2018.1449843.

[14] V. I. Bogachev, *Measure theory*, Springer Science & Business Media, 2007.

[15] A. Bonito and J. Pasciak, *Numerical approximation of fractional powers of elliptic operators*, Mathematics of Computation **84** (2015), no. 295, 2083–2110. doi:10.1090/S0025-5718-2015-02937-8.

[16] C. Buchheim, A. Caprara, and A. Lodi, *An effective branch-and-bound algorithm for convex quadratic integer programming*, Mathematical Programming **135** (2012), no. 1-2, 369–395. doi:10.1007/s10107-011-0475-x.

[17] C. Buchheim, R. Kuhlmann, and C. Meyer, *Combinatorial optimal control of semilinear elliptic PDEs*, Computational Optimization and Applications **70** (2018), no. 3, 641–675. doi:10.1007/s10589-018-9993-2.

[18] C. Carathéodory, *Über den Variabilitätsbereich der Koeffizienten von Potenzreihen, die gegebene Werte nicht annehmen*, Mathematische Annalen **64** (1907), no. 1, 95–115. doi:10.1007/BF01449883.

[19] L. Cesari, *Optimization — Theory and Applications*, Springer Verlag, 1983. doi:10.1007/978-1-4613-8165-5.

[20] P. G. Ciarlet, *Linear and nonlinear functional analysis with applications*, Vol. 130, Siam, 2013.

[21] C. Clason, F. Kruse, and K. Kunisch, *Total variation regularization of multi-material topology optimization*, ESAIM: Mathematical Modelling and Numerical Analysis **52** (2018), no. 1, 275–303. doi:10.1051/m2an/2017061.

[22] C. Clason and K. Kunisch, *Multi-bang control of elliptic systems*, Annales de l'Institut Henri Poincaré (c) Analyse Non Linéaire, 2014, pp. 1109–1130. doi:10.1016/j.anihpc.2013.08.005.

[23] ———, *A convex analysis approach to multi-material topology optimization*, ESAIM: Mathematical Modelling and Numerical Analysis **50** (2016), no. 6, 1917–1936. doi:10.1051/m2an/2016012.

[24] C. Clason, C. Tameling, and B. Wirth, *Vector-valued multibang control of differential equations*, SIAM Journal on Control and Optimization **56** (2018), no. 3, 2295–2326. doi:10.1137/16M1104998.

[25] R. J. Dakin, *A tree-search algorithm for mixed integer programming problems*, The Computer Journal **8** (1965), no. 3, 250–255. doi:10.1093/comjnl/8.3.250.

[26] R. Dautray and J.-L. Lions, *Mathematical Analysis and Numerical Methods for Science and Technology: Volume 5 Evolution problems I*, Springer-Verlag, 1992. doi:10.1007/978-3-642-58090-1.

[27] F. S. De Blasi and G. Pianigiani, *Evolution inclusions in non-separable Banach spaces*, Comment. Math. Univ. Carolin **40** (1999), no. 2, 227–250.

[28] K. Deckelnick and M. Hinze, *A note on the approximation of elliptic control problems with bang-bang controls*, Computational Optimization and Applications **51** (2012), no. 2, 931–939. doi:10.1007/s10589-010-9365-z.

[29] E. Di Nezza, G. Palatucci, and E. Valdinoci, *Hitchhiker's guide to the fractional Sobolev spaces*, Bulletin des Sciences Mathématiques **136** (2012), no. 5, 521–573. doi:10.1016/j.bulsci.2011.12.004.

[30] M. Diehl, *Real-time optimization for large scale nonlinear processes*, Ph.D. Thesis, Heidelberg University, 2001.

[31] J. Diestel and J. J. Uhl, *Vector measures*, American Mathematical Society, Providence, 1977. doi:10.1090/surv/015.

[32] N. Dinculeanu, *Vector measures*, VEB Deutscher Verlag der Wissenschaften, Berlin, 1967.

[33] K. Disser, A.F.M. ter Elst, and J. Rehberg, *On maximal parabolic regularity for non-autonomous parabolic operators*, Journal of Differential Equations **262** (2017), no. 3, 2039–2072. doi:10.1016/j.jde.2016.10.033.

[34] I. Dobrakov, *On representation of linear operators on* $C_0(T, X)$, Czechoslovak Mathematical Journal **21** (1971), no. 1, 13–30. `https://dml.cz/bitstream/handle/10338.dmlcz/101000/CzechMathJ_21-1971-1_3.pdf`.

[35] K.-J. Engel and R. Nagel, *One-parameter semigroups for linear evolution equations*, Vol. 194, Springer Science & Business Media, 1999. doi:10.1007/b97696.

[36] A. F. Filippov, *On some problems of optimal control theory*, Vestnik Moskowskovo Universiteta, Math **2** (1958), 25–32.

[37] H. Frankowska, *A priori estimates for operational differential inclusions*, Journal of Differential Equations **84** (1990), no. 1, 100–128. doi:10.1016/0022-0396(90)90129-D.

[38] A. Fügenschuh, B. Geißler, A. Martin, and A. Morsi, *The transport PDE and mixed-integer linear programming*, Dagstuhl seminar proceedings, 2009.

[39] R. V. Gamkrelidze, *On sliding optimal states*, Doklady akademii nauk, 1962, pp. 1243–1245.

[40] M. R. Garey and D. S. Johnson, *Computers and Intractability: A Guide to the Theory of NP-Completeness*, W.H. Freeman, New York, 1979. doi:10.1137/1024022.

[41] M. Gerdts, *Solving mixed-integer optimal control problems by Branch&Bound: A case study from automobile test-driving with gear shift*, Optimal Control Applications and Methods **26** (2005), 1–18. doi:10.1002/oca.751.

[42] ———, *A variable time transformation method for mixed-integer optimal control problems*, Optimal Control Applications and Methods **27** (2006), no. 3, 169–182. doi:10.1002/oca.778.

[43] M. Gerdts and S. Sager, *Mixed-Integer DAE Optimal Control Problems: Necessary conditions and bounds*, Control and Optimization with Differential-Algebraic Constraints, 2012, pp. 189–212.

[44] S. Göttlich, A. Potschka, and C. Teuber, *A partial outer convexification approach to control transmission lines*, Computational Optimization and Applications (2018Nov). doi:10.1007/s10589-018-0047-6.

[45] S. Göttlich, A. Potschka, and U. Ziegler, *Partial outer convexification for traffic light optimization in road networks*, SIAM Journal on Scientific Computing **39** (2017), no. 1, B53–B75. doi:10.1137/15M1048197.

[46] T. H. Grönwall, *Note on the derivatives with respect to a parameter of the solutions of a system of differential equations*, Annals of Mathematics (1919), 292–296. doi:10.2307/1967124.

[47] M. Gugat, G. Leugering, A. Martin, M. Schmidt, M. Sirvent, and D. Wintergerst, *MIP-based instantaneous control of mixed-integer PDE-constrained gas transport problems*, Computational Optimization and Applications **70** (2018), no. 1, 267–294. doi:10.1007/s10589-017-9970-1.

[48] M. Hahn and S. Sager, *Combinatorial integral approximation for mixed-integer pde-constrained optimization problems* (2018). ANL Preprint ANL/MCS-P9037-0118.

[49] F. Hante, *Stability and optimal control of switching PDE-dynamical systems*, arXiv preprint arXiv:1802.08143 (2018).

[50] F. M. Hante, *Relaxation methods for hyperbolic PDE mixed-integer optimal control problems*, Optimal Control Applications and Methods **38** (2017), no. 6, 1103–1110. doi:10.1002/oca.2315.

[51] F. M. Hante and S. Sager, *Relaxation methods for mixed-integer optimal control of partial differential equations*, Computational Optimization and Applications **55** (2013), no. 1, 197–225. doi:10.1007/s10589-012-9518-3.

[52] J. Haslinger and R. A. E. Mäkinen, *On a topology optimization problem governed by two-dimensional Helmholtz equation*, Computational Optimization and Applications **62** (2015), no. 2, 517–544. doi:10.1007/s10589-015-9746-4.

[53] D. Hilbert, *Über die stetige Abbildung einer Linie auf ein Flächenstück*, Mathematische Annalen **38** (1891), no. 3, 459–460. doi:10.1007/978-3-662-38452-7.

[54] M. Hintermüller and A. Laurain, *A shape and topology optimization technique for solving a class of linear complementarity problems in function space*, Computational Optimization and Applications **46** (2010), no. 3, 535–569. doi:10.1007/s10589-008-9201-x.

[55] M. Hinze, *A variational discretization concept in control constrained optimization: the linear-quadratic case*, Computational Optimization and Applications **30** (2005), no. 1, 45–61. doi:10.1007/s10589-005-4559-5.

[56] T. Hoheisel, *Mathematical programs with vanishing constraints*, Ph.D. Thesis, Würzburg University, 2009.

[57] T. Hoheisel and C. Kanzow, *First- and second-order optimality conditions for mathematical programs with vanishing constraints*, Applications of Mathematics **52** (2007), no. 6, 495–514. doi:10.1007/s10492-007-0029-y.

[58] ———, *On the Abadie and Guignard constraint qualifications for mathematical programmes with vanishing constraints*, Optimization **58** (2009), no. 4, 431–448. doi:10.1080/02331930701763405.

[59] T. Hoheisel, C. Kanzow, and J. V. Outrata, *Exact penalty results for mathematical programs with vanishing constraints*, Nonlinear Analysis: Theory, Methods & Applications **72** (2010), no. 5, 2514–2526. doi:10.1016/j.na.2009.10.047.

[60] T. Hytönen, J. Van Neerven, M. Veraar, and L. Weis, *Analysis in Banach spaces*, Springer, 2016. doi:10.1007/978-3-319-48520-1.

[61] K. Ito, K. Kunisch, and G. H. Peichl, *Variational approach to shape derivatives*, ESAIM: Control, Optimisation and Calculus of Variations **14** (2008), no. 3, 517–539. doi:10.1051/cocv:2008002.

[62] A. F. Izmailov and M. V. Solodov, *Mathematical programs with vanishing constraints: optimality conditions, sensitivity, and a relaxation method*, Journal of Optimization Theory and Applications **142** (2009), no. 3, 501–532. doi:10.1007/s10957-009-9517-4.

[63] E. Jones, T. Oliphant, P. Peterson, et al., *SciPy: Open source scientific tools for Python*, 2001. [Online; accessed October 28, 2019].

[64] M. Jung, *Relaxations and approximations for mixed-integer optimal control*, Ph.D. Thesis, Heidelberg University, 2013. doi:10.11588/heidok.00016036.

[65] M. N. Jung, G. Reinelt, and S. Sager, *The Lagrangian relaxation for the combinatorial integral approximation problem*, Optimization Methods and Software **30** (2015), no. 1, 54–80. doi:10.1080/10556788.2014.890196.

[66] C. Kirches, *Fast numerical methods for mixed-integer nonlinear model-predictive control*, Ph.D. Thesis, Heidelberg University, 2011.

[67] C. Kirches, H. G. Bock, J. P. Schlöder, and S. Sager, *Mixed-integer NMPC for predictive cruise control of heavy-duty trucks*, European Control Conference (ECC), 2013, pp. 4118–4123. doi:10.23919/ECC.2013.6669210.

[68] C. Kirches, F. Lenders, and P. Manns, *Approximation properties and tight bounds for constrained mixed-integer optimal control*, Optimization Online Preprint **5404** (2019). submitted.

[69] C. Kirches, S. Sager, H. G. Bock, and J. P. Schlöder, *Time-optimal control of automobile test drives with gear shifts*, Optimal Control Applications and Methods **31** (2010), no. 2, 137–153. doi:10.1002/oca.892.

[70] C. Kirchner, M. Herty, S. Göttlich, and A. Klar, *Optimal control for continuous supply network models*, Networks & Heterogeneous Media **1** (2006), no. 4, 675–688. doi:10.3934/nhm.2006.1.675.

[71] F. Lenders, *Numerical Methods for Mixed-Integer Optimal Control with Combinatorial Constraints*, Ph.D. Thesis, Heidelberg University, 2017.

[72] R. J. LeVeque, *Finite volume methods for hyperbolic problems*, Cambridge University Press, 2002. doi:10.1017/CBO9780511791253.

[73] J. Lindenstrauss, *A short proof of Liapounoff's convexity theorem*, Journal of Mathematics and Mechanics **15** (1966), no. 6, 971–972. doi:10.1512/iumj.1966.15.15064.

[74] J. Lindenstrauss and D. Preiss, *On Fréchet differentiability of Lipschitz maps between Banach spaces*, Annals of Mathematics (2003), 257–288. doi:10.2307/3597167.

[75] J.-L. Lions, *Quelques méthodes de résolution des problemes aux limites non linéaires*, Dunod, Gauthier-Villars, Paris, 1969.

[76] J.-L. Lions and E. Magenes, *Non-homogeneous boundary value problems, vol. 1*, Springer, 1972. doi:10.1007/978-3-642-65161-8.

[77] J. D. C. Little, K. G. Murty, D. W. Sweeney, and C. Karel, *An algorithm for the traveling salesman problem*, Operations Research **11** (1963), no. 6, 972–989. doi:10.1287/opre.11.6.972.

[78] A. A. Lyapunov, *On completely additive vector functions*, Izv. Akad. Nauk SSSR **4** (1940), 465–478.

[79] P. Manns and C. Kirches, *Multi-dimensional sum-up rounding for elliptic control systems*, submitted (2018). `https://spp1962.wias-berlin.de/preprints/080r.pdf`.

[80] ———, *Improved regularity assumptions for partial outer convexification of mixed-integer pde-constrained optimization problems*, ESAIM: Control, Optimisation and Calculus of Variations (2019). (to appear), doi:10.1051/cocv/2019016.

[81] ———, *Multi-dimensional Sum-Up Rounding using Hilbert curve iterates*, Proceedings in Applied Mathematics and Mechanics, 2019. (to appear), doi:10.1002/pamm.201900065.

[82] P. Manns, C. Kirches, and F. Lenders, *A linear bound on the integrality gap for sum-up rounding in the presence of vanishing constraints*, submitted (2017). `http://www.optimization-online.org/DB_HTML/2018/04/6580.html`.

[83] C. Marchal, *Chattering arcs and chattering controls*, Journal of Optimization Theory and Applications **11** (1973), no. 5, 441–468. doi:10.1007/BF00935659.

[84] A. Martin, M. Möller, and S. Moritz, *Mixed integer models for the stationary case of gas network optimization*, Mathematical Programming **105** (2006), no. 2-3, 563–582. doi:10.1007/s10107-005-0665-5.

[85] C. Meyer and L. M. Susu, *Optimal control of nonsmooth, semilinear parabolic equations*, SIAM Journal on Control and Optimization **55** (2017), no. 4, 2206–2234. doi:10.1137/15M1040426.

[86] R. H. Nochetto, E. Otárola, and A. J. Salgado, *A PDE approach to fractional diffusion in general domains: a priori error analysis*, Foundations of Computational Mathematics **15** (2015), no. 3, 733–791. doi:10.1007/s10208-014-9208-x.

[87] A. Pazy, *Semigroups of linear operators and applications to partial differential equations*, Vol. 44, Springer Science & Business Media, 1983. doi:10.1007/978-1-4612-5561-1.

[88] B. J. Pettis, *On integration in vector spaces*, Transactions of the American Mathematical Society **44** (1938), no. 2, 277–304. doi:10.1112/jlms/s1-31.4.399.

[89] M. E. Pfetsch, A. Fügenschuh, B. Geißler, N. Geißler, R. Gollmer, B. Hiller, J. Humpola, T. Koch, T. Lehmann, A. Martin, and others, *Validation of nominations in gas network optimization: models, methods, and solutions*, Optimization Methods and Software **30** (2015), no. 1, 15–53. doi:10.1080/10556788.2014.888426.

[90] O. Pironneau, *Optimal shape design for elliptic systems*, Springer, 1984. doi:10.1007/978-3-642-87722-3.

[91] R. Pluciennik, *On some properties of the superposition operator in generalized Orlicz spaces of vector-valued functions*, Comment. Math. Parce Mat. **25** (1985), 321–337.

[92] H. Rademacher, *Über partielle und totale Differenzierbarkeit von Funktionen mehrerer Variabeln und über die Transformation der Doppelintegrale*, Mathematische Annalen **79** (1919), no. 4, 340–359. doi:10.1007/BF01498415.

[93] J. P. Raymond and H. Zidani, *Hamiltonian Pontryagin's principles for control problems governed by semilinear parabolic equations*, Applied Mathematics and Optimization **39** (1999), no. 2, 143–177. doi:10.1007/s002459900102.

[94] M. Renardy and R. C. Rogers, *An Introduction to Partial Differential Equations*, Vol. 13, Springer Science & Business Media, 2006. doi:10.1007/b97427.

[95] B. Russell, *The Scientific Outlook*, 2nd ed., George Allen & Unwin Ltd., 1949. doi:10.5840/monist193343225.

[96] E. W. Sachs and S. Volkwein, *POD-Galerkin approximations in PDE-constrained optimization*, GAMM-Mitteilungen **33** (2010), no. 2, 194–208. doi:10.1002/gamm.201010015.

[97] S. Sager, *Numerical methods for mixed-integer optimal control problems*, Der andere Verlag Tönning, Lübeck, Marburg, 2005.

[98] ———, *Reformulations and Algorithms for the Optimization of Switching Decisions in Nonlinear Optimal Control*, Journal of Process Control **19** (2009), no. 8, 1238–1247. doi:10.1016/j.jprocont.2009.03.008.

[99] ———, *Sampling decisions in optimum experimental design in the light of Pontryagin's maximum principle*, SIAM Journal on Control and Optimization **51** (2013), no. 4, 3181–3207. doi:10.1137/110835098.

[100] S. Sager, H. G. Bock, and M. Diehl, *The Integer Approximation Error in Mixed-Integer Optimal Control*, Mathematical Programming, Series A **133** (2012), no. 1–2, 1–23. doi:10.1007/s10107-010-0405-3.

[101] S. Sager, M. Jung, and C. Kirches, *Combinatorial Integral Approximation*, Mathematical Methods of Operations Research **73** (2011), no. 3, 363–380. doi:10.1007/s00186-011-0355-4.

[102] S. Sager, C. Kirches, and H. G. Bock, *Fast solution of periodic optimal control problems in automobile test-driving with gear shifts*, 47th IEEE Conference on Decision and Control, 2008, pp. 1563–1568. doi:10.1109/CDC.2008.4739014.

[103] K. Schmüdgen, *Unbounded self-adjoint operators on Hilbert space*, Vol. 265, Springer Science & Business Media, 2012. doi:10.1007/978-94-007-4753-1.

[104] B. Schweizer, *Partielle Differentialgleichungen: Eine anwendungsorientierte Einführung*, Springer-Verlag, 2013. doi:10.1007/978-3-642-40638-6.

[105] M. Siebenborn, *A shape optimization algorithm for interface identification allowing topological changes*, Journal of Optimization Theory and Applications **177** (2018), no. 2, 306–328. doi:10.1007/s10957-018-1279-4.

[106] O. Sigmund and K. Maute, *Topology optimization approaches*, Structural and Multidisciplinary Optimization **48** (2013), no. 6, 1031–1055. doi:10.1007/s00158-013-0978-6.

[107] B. Simon, *Operator theory. a comprehensive course in analysis, part 4*, American Mathematical Society, Providence (2015). doi:10.1090/simon/004.

[108] J. Simon, *Compact sets in the space* $L^p((0,T),B)$, Annali di Matematica pura ed applicata **146** (1986), no. 1, 65–96. doi:10.1007/BF01762360.

[109] G. Stadler, *Elliptic optimal control problems with* L^1*-control cost and applications for the placement of control devices*, Computational Optimization and Applications **44** (2009), no. 2, 159. doi:10.1007/s10589-007-9150-9.

[110] E. M. Stein, *Singular integrals and differentiability properties of functions*, Princeton Mathematical Series, vol. 30, Princeton University Press, 1970. http://www.jstor.org/stable/j.ctt1bpmb07.

[111] E. M. Stein and R. Shakarchi, *Real analysis: measure theory, integration, and Hilbert spaces*, Princeton University Press, 2005. doi:10.1017/S0025557200181343.

[112] M. C. Steinbach, *On PDE solution in transient optimization of gas networks*, Journal of Computational and Applied Mathematics **203** (2007), no. 2, 345–361. doi:10.1016/j.cam.2006.04.018.

[113] R. W. Steinberg and L. Floyd, *An adaptive algorithm for spatial greyscale*, Proceedings of the Society of Information Display **17** (1976), 75–77.

[114] L. Tartar, *Compensated compactness and applications to partial differential equations*, Nonlinear analysis and mechanics: Heriot-Watt symposium, 1979, pp. 136–212.

[115] ———, *An introduction to Sobolev spaces and interpolation spaces*, Vol. 3, Springer Science & Business Media, 2007. doi:10.1007/978-3-540-71483-5.

[116] G. Teschl, *Ordinary differential equations and dynamical systems*, Vol. 140, American Mathematical Society Providence, 2012. doi:10.1090/gsm/140.

[117] G. Wachsmuth and D. Wachsmuth, *Convergence and regularization results for optimal control problems with sparsity functional*, ESAIM: Control, Optimisation and Calculus of Variations **17** (2011), no. 3, 858–886. doi:10.1051/cocv/2010027.

[118] J. Warga, *Necessary conditions for minimum in relaxed variational problems*, Journal of Mathematical Analysis and Applications **4** (1962), no. 1, 129–145. doi:10.1016/0022-247X(62)90034-3.

[119] ———, *Necessary conditions without differentiability assumptions in optimal control*, Journal of Differential Equations **18** (1975), no. 1, 41–62. doi:10.1016/0022-0396(75)90080-7.

[120] T. Ważewski, *On an optimal control problem*, Differential Equations and their Applications (1963), 229–242. doi:10338.dmlcz/702189.

[121] J. A. Yorke, *Another proof of the Liapunov convexity theorem*, SIAM Journal on Control **9** (1971), no. 3, 351–353. doi:10.1137/0309025.

[122] V. M. Zavala, J. Wang, S. Leyffer, E. M. Constantinescu, M. Anitescu, and G. Conzelmann, *Proactive energy management for next-generation building systems*, Proceedings of SimBuild **4** (2010), no. 1, 377–385.

[123] C. Zeile, T. Weber, and S. Sager, *Combinatorial integral approximation decompositions for mixed-integer optimal control*, submitted (2018). `http://www.optimization-online.org/DB_FILE/2018/02/6472.pdf`.

Acronyms

ACP Abstract Cauchy Problem
BV Bounded Variation
BVP Boundary Value Problem
CIA Combinatorial Integral Approximation
DAE Differential-Algebraic Equation
ESRC Elliptic Subproblem in (RC)
FTC Fundamental Theorem of Calculus
IP Integer Program
IVP Initial Value Problem
IBVP Initial Boundary Value Problem
MILP Mixed-Integer Linear Program
MINLP Mixed-Integer Nonlinear Program
MIOCP Mixed-Integer Optimal Control Problem
MIPDECO Mixed-Integer PDE-Constrained Optimization Problem
MPEC Mathematical Program with Complementarity Constraints
MPVC Mathematical Program with Vanishing Constraints
NFR Next-Forced Rounding
OCP Optimal Control Problem
ODE Ordinary Differential Equation
PDE Partial Differential Equation
RNP Radon-Nikodym Property
SUR Sum-Up Rounding
SOS1 Special Ordered Set of Type 1
VOC Variation of Constants Formula

Acknowledgments

The author acknowledges funding by Deutsche Forschungsgemeinschaft through Priority Programme 1962 under grant agreement KI1839/1-1.

During the preparation of this thesis my supervisor Christian Kirches provided very generous support in many ways. I highlight his exemplary *open door policy* and his high availability for mathematical discussions. All of it has been greatly appreciated and I like to thank for it. Furthermore, I like to thank Dirk Lorenz for sharing his remarkable knowledge about analysis in many discussions. I also like to thank Christian Meyer for his hospitality and the opportunity to spend a very fruitful week at his research group in Dortmund, which inspired several parts of this work (and where I also watched a match of the great BVB!). Many others contributed to this work in many ways and although I cannot do them justice here, I like to mention the following people in alphabetical order: Gerhard Baur, Lilli Bergner, Felix Bestehorn, Florian Bürgel, Mathis Fricke, Jan Glaubitz, Christoph Hansknecht, Robert Haller-Dintelmann, Daniel Hauer, Imke Joormann, Birgit Komander, Manuel Kudruss, Alexander Kreiß, Felix Lenders, Rebecca Nahme, Cornelia Pioch, Andreas Potschka, Hendrik Ranocha, Pascal Richter, Sebastian Stiller, Stefan Ulbrich, Silke Thiel and Ingo von Laer. Last but not least, I like to thank Silvi Ewald for her constant support and courage regarding the choices of life we made on the way to this work.